# Quantum Machine Learning in Chemical Space

**Inauguraldissertation**
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelgt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität

von

Felix Andreas Faber

aus Schweden

2019

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Anatole von Lilienfeld, and Prof. Dr. Stefan Goedecker

Basel, 25.06.2019

Prof. Dr. Martin Spiess
The Dean of Faculty

## Abstract

This thesis focus on the overlap of first principle quantum methods and machine learning in computational chemistry and materials science, commonly referred to as Quantum Machine Learning (QML).

Assessing and benchmarking the performance of existing machine learning models on various classes of compounds and chemical properties is a substantial part of this thesis. These results are used to understand better which machine learning models are best suited for a given combination of properties and compounds. For example, thirteen electronic ground state properties of ∼131k organic molecules, calculated at hybrid-DFT level of theory, were used to gauge the predictive accuracy of combinations of representations and regressors. The out-of-sample prediction errors of the models on the hybrid-DFT quality data are on par with, or close to, the CCSD(T) error to experimental values, indicating that reference data need to go beyond hybrid-DFT if QML predictions are to surpass chemical accuracies.

Another area of focus is on developing new and accurate QML models. A new representation of atoms in its chemical environment is introduced, by rethinking the way structural and chemical compound information is encoded into training data. The representation interpolates elemental properties across both atoms and compounds, making it well suited for datasets with high compositional and structural degrees of freedom. Numerical results evidence that, compared to current benchmarks, this representation yield superior predictive power in combination with kernel ridge regression on a diverse set of systems, including diverse organic molecules, non-covalently bonded protein side-chains, water clusters, and crystalline solids. Furthermore, the role of response operators when learning response properties of the energy is discussed, leading to a formalism for learning response properties of the energy by applying the corresponding response operator directly to the quantum machine learning model. Using this formalism leads to train QML models results in lower out-of-sample errors than learning the corresponding properties directly. The formalism can also be used to reproduce accurate normal modes and IR-spectra in molecules.

Finally, the applicability of QML models is explored. A machine learning model which encodes the elemental identities of the atoms placed in each site, to exhaustively screen the formation energy of ∼2 milion Elpasolite crystals. The resulting model's accuracy improves systematically with additional training data, reaching an accuracy of 0.1 eV/atom when trained on 10 k crystals. Out of the ∼ 2 million crystals, we identify 90 unique structures which span the

convex hull of stability, among which $NFAl_2Ca_6$, with uncommon stoichiometry and a negative atomic oxidation state for Al.

## Publications

1. F. A. Faber, A. S. Christensen, O.A. von Lilienfeld, "Quantum Mechanical Response Operators and Machine Learning", *Chapter in "Machine Learning for Quantum Simulations of Molecules and Materials", under review*, (2019)

2. F. A. Faber, O.A. von Lilienfeld, "Modeling Materials Quantum Properties with Machine Learning", *Chapter in "Materials Informatics", in press* (2019)

3. A. S. Christensen, F. A. Faber, O.A. von Lilienfeld, "Operators in Machine Learning: Response Properties in Chemical Space", *J. chem. Phys.*, (2019)

4. F. A. Faber, A. S. Christensen, B. Huang, O.A. von Lilienfeld, "Alchemical and Structural Distribution Based Representation for Universal Quantum Machine Learning", *J. chem. Phys.*, (2018)

5. F. A. Faber et al., "Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error", *J. Chem. Theory Comput*, (2017)

6. F. A. Faber, A. Lindmaa, O.A. von Lilienfeld and R. Armiento, "Machine Learning Energies of 2 Million Elpasolite ($ABC_2D_6$) Crystals", *Phys. Rev. Lett.* 117: 135502, (2016)

7. F. Faber, A. Lindmaa, O.A. von Lilienfeld and R. Armiento, "Crystal Structure Representations for Machine Learning Models of Formation Energies", *Int. J. Quant. Chem.* 115: 1094, (2015)

## Acknowledgements

I would like to express my sincerest gratitude towards my supervisor and mentor, Anatole von Lilienfeld. Anatole supported and guided me throughout my PhD studies. He taught me how to conduct good research, and how to carry myself as a scientist.

I am also thankful to Patrick Riley and the other collaborators at Google for allowing me the opportunity to work with them. Anders Steen Christensen and Rickard Armiento worked closely with me on several projects, and I would like to thank them for their help and friendship. Samuel Mossberg deserves my gratitude for helping me to verify some of my mathematical derivations.

I would also like to thank Prof. Stefan Goedecker for agreeing to co-examine my thesis.

Anders Steen Christensen, Jimmy Kromann, and Morris Trestman proofread my thesis and I am very grateful for their help.

I would also like to thank Diana Tahchieva, Stefan Heinen, Max Schwilk, Pál Mezei, Guido von Rudorff, Bing Huang, Marco Bragato, Bernard Mazouin, Jakub Wagner, Dirk Bakowies, Amancaya Formica and all other group members and colleagues, past and present, whom I have worked with during my P.hD.

Finally, I wish to thank my family and friends from Sweden, as well as the new friends that I've made during my PhD, all whom been very supportive.

## Abbreviations

**ADF** Angular Distribution Function

**BAML** Bonds, Angles, Machine Learning

**CV** Cross Validation

**DFT** Density Functional Theory

**ECFP4** Extended Connectivity Fingerprint

**EN** Elastic Net

**FCHL** Faber, Christensen, Huang, Lilienfeld

**FLLA** Faber, Lindmaa, Lilienfeld, Armiento

**GDML** Gradient Domain Learning

**GPR** Gaussian Process Regression

**HD** Histogram of Distances

**HDA** Histogram of Distances and Angles

**HDAD** Histogram of Distances, Angles, and Dihedral Angels

**HOMO** Highest Occupied Molecular Orbital

**ICSD** Inorganic Crystal Structure Database

**IR** Infra Red

**KRR** Kernel Ridge Regression

**LC** Learning Curve

**LUMO** Lowest Unoccupied Molecular Orbital

**MARAD** Molecular atomic radial angular distribution

**MBTR** Many-Body Tensor Representation

**MG** Molecular Graphs

**ML** Machine Learning

**MP** Materials Project

**MPD** Materials Project Dataset

**NN** Neural Network

**OQMD** Open Quantum Materials Database

**OQML** Operator Quantum Machine Learning

**QM** Quantum Mechanical

**QML** Quantum Machine Learning

**RDF** Radial Distribution Function

**RF** Random Forest

**RMSE** Root Mean Squared Error

**SLATM** Spectral London Axilrod-Teller-Muto

**SOAP** Smooth Overlap of Atomic Potentials

**SVD** Singular Value Decomposition

**UFF** Universal Force Field

**ZPVE** Zero Point Vibrational Energy

# Chapter 1

# Introduction

Since the 1940s, computational simulations have been used to help understand chemical and material sciences. If done correctly, they can provide insight into why a chemical reaction occurs [1, 2], identify suitable drug candidates for diseases [3–5], and help to discover new materials with exotic properties [6–8].

While experiments will always be essential tools, computer simulations pose several advantages. In comparison to experiments, computational simulations allow systematic control of all relevant variables. Simulations therefore produce deterministic results without statistical noise. Furthermore, a computer simulation is often easier to set up and cheaper to run than its experimental counterpart.

However, calculating a given system property is generally a trade-off between accuracy and computational speed. For example, quantum mechanical methods exist that produce results which closely match experimental values, but whose computational cost grows rapidly with system size and complexity. These methods include density functional theory (DFT) [9, 10], post-Hartree Fock methods [11–15] and quantum Monte Carlo [16], which approximate solutions to the electronic Schrodinger equation; they therefore provide consistent estimations without extensive parametrisation at calculation times that can reach into weeks, or more.

On the other hand, force-fields and coarse-grained models can calculate properties of larger systems on the timescale of milliseconds. However, their predetermined functional forms limit their applicability to specific problem sets and many force-fields struggles with bond-breaking. Furthermore, developing new force-fields is notoriously difficult and extensive parameterization is necessary to re-task them to areas outside their intended design.

The nascent field of machine learning (ML) poses a different approach to this speed/accuracy trade-off by predicting system properties, instead of direct calculation. This prediction

arises through inference from compounds where the properties are known. In contrast to the force-fields and coarse-grained methods, the prediction error of a ML model tends to decrease systematically with the number of compounds used to fit the model. Hence, given enough examples, a ML model can in principle reach arbitrary predictive accuracies.

Because Machine learning models are inductive, their predictions can only be as accurate as their training data; therefore high quality reference data is needed. Ideally, the reference data should be free of statistical noise, arbitrarily accurate, and easily produced. Data produced by quantum mechanical methods satisfy all of these criteria.

This overlap of quantum mechanical methods and machine learning, known as Quantum Machine Learning (QML), forms the body of this thesis.

While other established scientific disciplines such as bio- and cheminformatics already use ML models trained on coarse-grained properties, such as toxicities and binding affinities, QML is constructed around fundamental quantum mechanical properties, such as energies and forces. The implication of such elementary focus, as highlighted in Ref. [17], is that QML can in principle be used to predict properties of all systems throughout chemical space.

## 1.1 Overview

This thesis can be summarized in three categories: (i) Developing QML models with applications in chemistry and the material sciences in order to predict chemical properties such as energies, forces or dipole moments from compound structures, thereby accelerating computational speed and improving prediction accuracies. Further discussed from chapter 4 to 7. (ii) Exploring the applicability of QML models in various scenarios, as discussed in chapters 6 and 7. (iii) Assessing and benchmarking the performance of existing QML models on various classes of compounds and chemical properties to better identify the best model and property combinations. Discussed further in chapters 4, 5 and 6.

The thesis is organized as follows:

Chapter 2 summarizes the underlying theory of QML in the context of quantum chemistry. Furthermore, it provides an overview of the current state of the field, and the advancements made over the last decade.

Chapter 3 briefly discusses the meaning of chemical space and describes the datasets used for training and testing the QML models.

Chapter 4 contains a comprehensive benchmarking of the predictive performance of several QML model property combinations, as published in Ref. [18]. This study was done in collaboration with the Google Accelerated Science Group.

Chapter 5 discusses a new representation, based on atomic densities and elemental smearing. This study was published in Ref. [19].

Chapter 6 introduces a method for learning quantum response properties with the help of response operators, as published in Ref. [20].

Chapter 7 discusses the use of QML to predict the energies of $\sim$ 2 million Elpasolite crystal structures, as published in Ref. [8].

Chapter 8 summarizes and concludes the thesis.

# Chapter 2

# Quantum Machine Learning

The pioneer of ML, Author Samuel, characterizes ML as "the field of study that gives computers the ability to learn without being explicitly programmed" [21]. One could program a computer to play a board game by providing the computer with a set of instructions, depending on the current state of the board. While this approach is theoretically possible, it is often technically unfeasible because of the complexity of the task. Furthermore, even a well-designed set of decisions will be static, unable to respond to unforeseen strategies or, like a human player, improve as the game progresses. Rather than devise an algorithm with a fixed set of decisions, Samuel's idea was to let the computer improve itself after every game. While the initial performance of the computer was poor, it soon improved dramatically by playing thousands of games against itself, and easily defeated Samuel.

In QML, a ML model $f(C, \alpha)$ is trained on a set compounds $\mathbf{C}^{\text{train}}$ with associated properties $\mathbf{p}^{\text{train}}$, as seen in Fig. 0.1.

A QML model $f(C, \alpha)$ trained on a training set that contains compounds $\mathbf{C}^{\text{train}}$ with already known properties $p$ can be used to infer the property $p_q$ of an unseen query compound $C_q$.

This chapter provides a brief introduction to ML techniques and practices in the context of quantum chemistry. Kernel ridge regression (KRR) and neural networks (NN) are the most widespread ML modes in QML and will therefore be covered. A brief discussion follows on how to evaluate ML models and how to best represent a compound to a ML model. The chapter ends with a summary of the current state of the field and an outline of some areas of QML currently undergoing active research.

## 2.1 Kernel Ridge Regression

KRR [22–25] and other kernel based models are among the most commonly used ML models

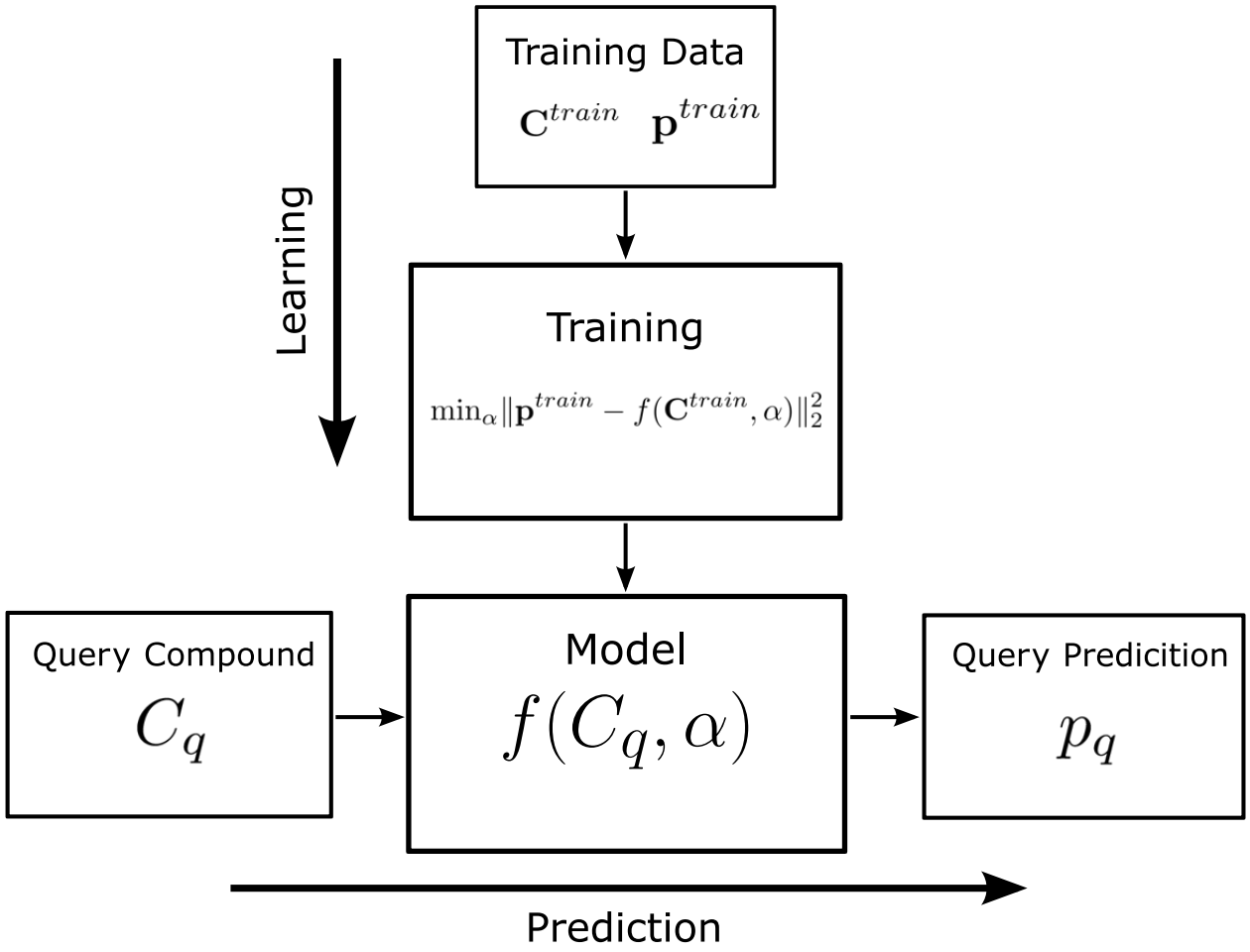



**Figure 0.1:** Flowchart depicting the training and prediction of a QML model. Horizontal axis depicts how a QML model predicts a property $p_q$ of a query compound $C_q$. Vertical axis shows how the QML model is trained on existing data. In this case it is done by finding the $\alpha$ that minimizes the euclidean distance between the properties $\mathbf{p}^{train}$ and the QML model $f(C, \alpha)$ for all compounds $\mathrm{C}^{train}$ in the training set.

within molecular and material science [26–33]. It is also the most used ML model throughout this thesis.

A property $p_q$ of a query compound $C_q$ is predicted by a sum of weighted kernels $K(C_q, C_i^{train})$ between $C_q$ and all compounds $C_i^{train}$ in the training set.

$$p(C_q) = \sum_i^N \alpha_i K(C_q, C_i^{train}) \tag{1.1}$$

The model is optimized by finding the set of $\alpha$s that minimizes the Euclidean distance between the model and the target property of the compounds in the training set, seen in eq. 1.2. Large values of $\alpha_i$ tend to be linked to overfitting, causing high generalization errors. It is therefore common practice to add Tikhonov regularization [34, 35], which penalize large $\alpha$.

$$\mathcal{J}(\alpha) = \min_\alpha ||\mathbf{p}_{ref}^{train} - \mathbf{p}_{est}^{train}||_2^2 + \lambda \alpha^t \mathbf{K} \alpha \tag{1.2}$$

$$= \min_\alpha (\mathbf{p} - \mathbf{K}\alpha)^t (\mathbf{p} - \mathbf{K}\alpha) + \lambda \alpha^t \mathbf{K} \alpha \tag{1.3}$$

The elements of the kernel matrix $\mathbf{K}$ correspond to the kernel function of the elements in the training set: $\mathbf{K}_{i,j} = K(C_i^{train}, C_j^{train})$. $\lambda$ is used to adjust the strength of the Tikhonov regularization. In practice, however, $\lambda$ can be set close to zero, as calculated training data is virtually free from statistical noise. Setting $\lambda$ somewhere between $10^{-6}$ and $10^{-12}$ is generally enough to ensure invertibility (due to finite numerical precision), and larger values tend to lower the performance of the model.

Equation 1.2 poses a convex minimization problem and is therefore equivalent to finding the solution which is stationary with respect to $\alpha$ , as seen in eq. 1.4.

$$\begin{aligned} \frac{\partial}{\partial \alpha}((\mathbf{p} - \mathbf{K}\alpha)^t(\mathbf{p} - \mathbf{K}\alpha) + \lambda \alpha^t \alpha) = \mathbf{0} \Leftrightarrow \\ 2\mathbf{p}^t\mathbf{K} + 2\alpha^t\mathbf{K}^2 + 2\lambda\alpha^t\mathbf{K} = \mathbf{0} \Leftrightarrow \\ \mathbf{p}\mathbf{K} = \alpha(\mathbf{K} + \lambda\mathbf{I})\mathbf{K} \end{aligned} \tag{1.4}$$

$\mathbf{K}$ is positive-definite, so the $\alpha$ which minimizes eq. 1.2 can be obtained using eq. 1.5.

$$\mathbf{p}(\mathbf{K} + \lambda\mathbf{I})^{-1} = \alpha \tag{1.5}$$

The kernel function is a positive definite symmetric function. This means that a matrix whose elements consist of pairwise evaluations between all training samples $\mathbf{K}_{i,j} = K(C_i^{train}, C_j^{train})$ only has positive eigenvalues. An alternative way of viewing the kernel function is as the basis of the regression model. A kernel function is placed on each training instance, dynamically growing the flexibility of the model with training set size.

Commonly used kernel functions include linear, Gaussian and Laplacian, seen in eq. 1.6, where $\sigma$ is a hyper-parameter, adjusting the width of the kernel.

The hyper-parameters need to be optimized separately from the regression coefficients using a logarithmic grid-search, or more sophisticated heuristics. Furthermore, any linear combination of kernel functions is a valid kernel function, and the best choice of kernel function should be selected for the problem at hand.

$$\text{Linear: } K(C, C') = C^t C' \tag{1.6}$$

$$\text{Gaussian: } K(C, C') = \exp(-\frac{||C - C'||_2^2}{2\sigma^2}) \tag{1.7}$$

$$\text{Lapacian: } K(C, C') = \exp(-\frac{||C - C'||_1}{\sigma}) \tag{1.8}$$

## 2.2 Neural Networks

NN have in the past decade proven to be exceptionally well suited for solving complicated classification problems. For example, state-of-the-art networks can classify 12 million images with over 20000 classes correctly more then 96% of the time [36, 37]. NN models have also shown promise in QML in recent years [18, 30, 38–45].

The simplest form of a NN consists of nodes stacked in layers, as seen in Fig. 2.2, where each node in a layer is a nonlinear transformation of the nodes in the previous layer. The first layer passes the representation of a query compound $C_q$, in the form of a vector, to the first hidden layer $\mathbf{x}^1$ in the network. This is followed by $L$ hidden layers, where the output $\mathbf{x}^{i-1}$ from a previous layer $i-1$ is used as input for the next layer $i$. The connections between layers consist of an affine transformation followed by an entrywise non-linear transformation through an activation function $\phi$.

The affine transformation consists of a matrix multiplication $\mathbf{w}$ with the vector from the previous layer $\mathbf{x}^{\text{previous}}$ and a additive bias $\mathbf{b}$, seen in Eq. 2.9.

$$\mathbf{W}\mathbf{x}^{\text{previous}} - \mathbf{b} \tag{2.9}$$

The activation function is necessary because no new information would be introduced beyond the first layer if only affine transformations connected the layers. Some commonly used activation functions are sigmoid, tanh and ReLu seen in Eq. 2.10.

$$\text{sigmoid} : \frac{1}{1+e^{-x}} \tag{2.10}$$

$$\tanh : \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.11}$$

$$\text{ReLu} : \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \tag{2.12}$$

$$\tag{2.13}$$

Eq. 2.14 describes the connection from the query compound $C_q$ to the first hidden layer $\mathbf{x}^1$. Eq. 2.15 describes the connection from the $i-1$'th layer to $i$'th layer. The activation function is generally omitted in the final layer $\mathbf{x}^L$, which is mapped on the target property $p_q$, as seen in 2.16.

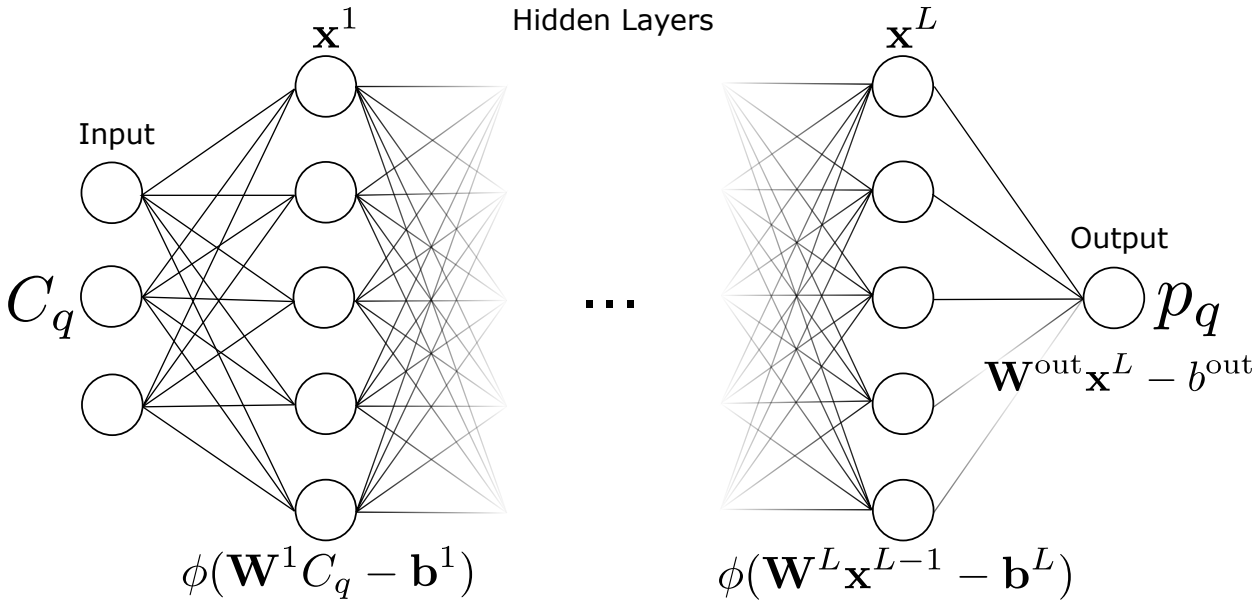



**Figure 2.2:** Schematic representation of a NN consisting of $L$ fully connected hidden layers. The neural network takes a input vector $C_q$ and predicts an output property $p_q$

$$\mathbf{x}^1 = \phi(\mathbf{W}^1 C_q - \mathbf{b}^1) \tag{2.14}$$

$$\mathbf{x}^i = \phi(\mathbf{W}^i \mathbf{x}^{i-1} - \mathbf{b}^i) \tag{2.15}$$

$$p_q = \mathbf{W}^{\text{out}} \mathbf{x}^L - \mathbf{b}^{\text{out}} \tag{2.16}$$

The hidden layers in the architecture above are called a fully connected layer. In practice, however, most NNs utilize many types of connections between layers, such as convolutional layers [46], recurrent layers [47], or residual layers [37].

Finding the optimal regression coefficients in a NN is a non-linear problem and has to be obtained numerically. However, calculating the gradient with respect to the regression coefficients using finite difference would require roughly one function evaluation for each parameter in the NN. It is not uncommon for a NN to have millions of regression coefficients, meaning that it just one gradient step quickly becomes prohibitively expensive.

The gradients are instead obtained using an iterative algorithm, called Backpropagation [48]. Backpropagation is, in essence, a clever use of the chain-rule to obtain the gradient at the cost of roughly two function evaluations.

Nowadays, NNs are often trained using more sophisticated gradient-based optimization methods, such as stochastic gradient decent [49], Limited memory BFGS [50] or ADAM [51].

## 2.3 Learning Curves

Typically the out-of-sample error $\epsilon$ follows a power-law decay with respect to the number of training samples $N_{train}$, as seen in Eq. 3.17 [25, 52, 53].

$$\epsilon \propto bN^{-a} \tag{3.17}$$

A plot of $\epsilon$ against $N_{train}$ on a log-log scale, as shown in Fig. 4.3, should therefore decrease linearly. The exponent $a$ and the logarithm of the prefactor $b$ in Eq. 3.17 correspond to the slope and offset of the curve, respectively.

The learning curve (LC) can be used to assess the aptitude of a model. For example, as seen in Fig. 4.3, a good model generally decays linearly for larger training set sizes. On the other hand, the learning rate of an inferior ML model diminishes with increasing training set size, preventing the model from reaching arbitrary target accuracies. Furthermore, a LC which is linear in a log-log scale can be easily extrapolated to give an estimate of how much data is needed to reach a target accuracy, as seen in Fig. 4.3.

## 2.4 Cross-Validation

ML models are typically most accurate on the data used to parametrize the model. However, generalizability and accurate predictions of new data are desirable; it is therefore reasonable to withhold parts of the data during training and later use it to evaluate the model's accuracy. A common practice is to split up the dataset into three parts: A training set, a validation set, and a test set.

The training set is used to fit the model to the target function. This can, for example, be done by finding the optimal regression coefficients $\alpha$, $\mathbf{W}$ and $\mathbf{b}$. The validation set is used to tweak the model. The tweaking includes optimizing hyperparameters, choosing a suitable kernel function or finding the best NN architecture. Finally, the performance of the model is evaluated on the test set, resulting in an estimate of how well the model generalizes to new compounds.

The above is one of the most simple cross-validation (CV) schemes. The next two sections will discuss k-fold CV and Random sub-sampling which tend to be more robust, especially for small datasets.

k-fold CV [54] is one of the most commonly used CV methods. The dataset is randomly split into $k$ subsets of roughly equal size. $k-1$ of the subsets are then used to train the model and the last subset is used to validate it. This process is repeated $k$ times, with each subset used once as the validation set. The errors obtained from the $k$ folds are then averaged. Typical choices of $k$ values are 5 or 10 [55].

This CV method ensures that every training sample is used both for training and validation.

In random sub-sampling cross-validation [54], the dataset is randomly split into several training and validation sets. The model is then trained and tested on each training and validation set, respectively, followed by an averaging of the error from all splits.

Random sub-sampling suffers from the possibility of overlooked data - due to the randomness

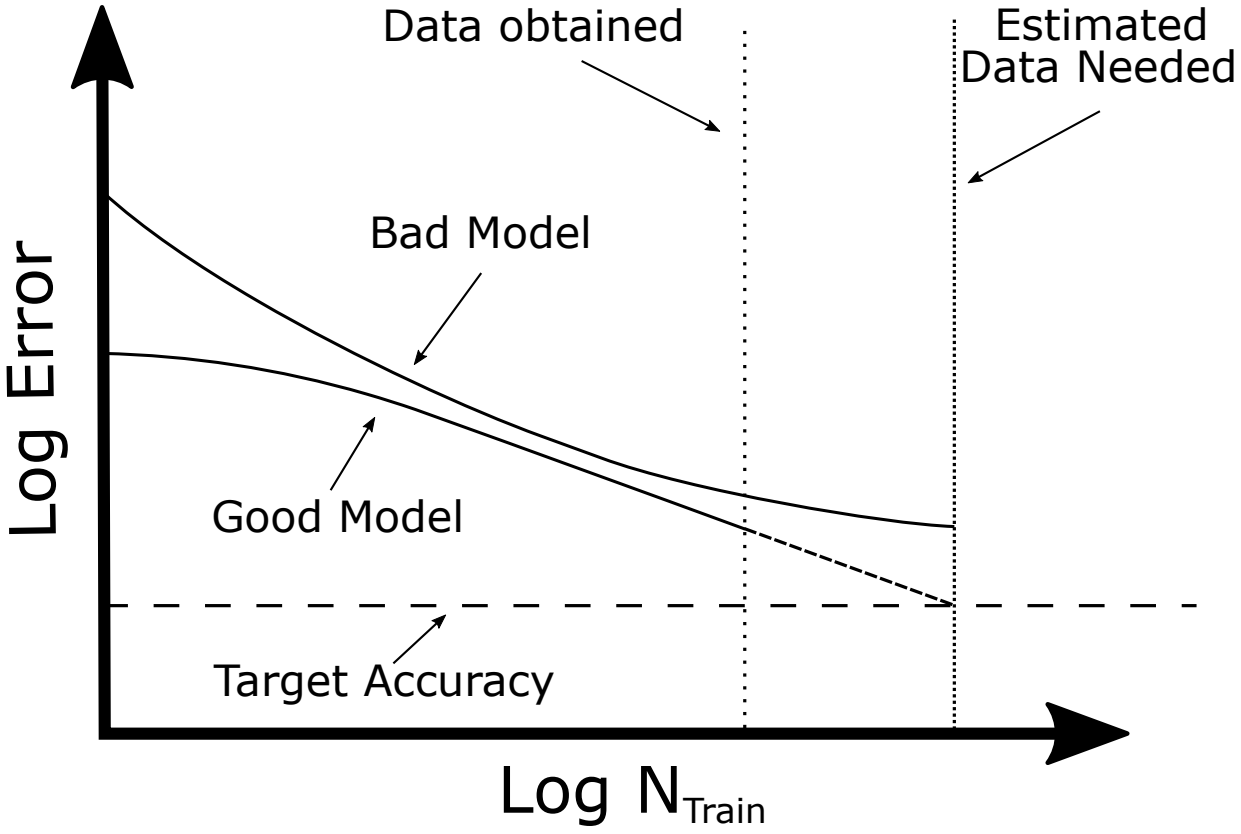



**Figure 4.3:** Schematic representation of typical LCs from good and bad ML models. The vertical and horizontal axes respectively represent the error and training set size on a logarithmic scale. The LC of a good ML model exhibits linear decay when plotted on a logarithmic scale for a large $N_{\text{train}}$. A less favorable model's decay rate slows down with increasing training set size, leading to stagnant learning. Furthermore, as seen in the figure, the training data needed to reach a desired target accuracy can easily be estimated for a ML model whose LC decreases linearly.

of its allocations, some training samples might never be included in the validation set, and vice versa.

However, random sub-sampling has the advantage that the proportion between the training and validation set sizes can be chosen at will, and is independent from the number of folds used. This differs from k-fold CV, whose training/validation set size proportion is necessarily dictated by the number of folds. Such customization makes random sub-sampling CV suitable for generating LCs, discussed in the following section, where the training/validation ratio is treated as a variable.

## 2.5 Representations

A representation is an aggregated feature which somehow encodes the composition and structure of a compound $C$, usually in the form of a vector. QML models use a representation input, which plays a significant role in the predictive accuracy of the entire ML model. This section will discuss the attributes seen in the best-performing representations.

An injective mapping between the compound and representation should exist. This means that any two compounds $C_1$ and $C_2$ are mapped to two distinct representations unless $C_1 = C_2$. This injective mapping is crucial because the representation will otherwise fail to distinguish between compounds, which hampers the resulting ML model's ability to make accurate predictions. The effects of non-injective representations can often be observed in their resulting LCs, which typically flatten out, as seen in Fig. 4.3.

An example of a non-injective mapping between compound and representation is one consisting only of unordered atom-atom pairwise features. As mentioned in Hansen *et al.* [30], and later discussed further and exemplified by Huang and von Lilienfeld [56], such non-injective representations fail to distinguish homeometric molecules.

Another example of a non-injective representation is one that is based purely of molecular graphs [57–59]. These representations will not be able to distinguish between different conformers.

Additionally, a representation should encode both scalar valued and tensorial properties. While scalar properties are invariant to both rotations and translations, tensorial properties such as forces, dipole moments, and polarizability should rotate with the input compound, i.e., they are covariant. One of the numerous ways of incorporating these symmetries into a ML model is to make the kernel/representation covariant [28, 60, 61].

Another way to circumvent the problem of learning tensorial and vector properties is to use a surrogate model. The surrogate model predicts scalar-properties, which are then used to generate the desired tensorial properties. For example, Gastegger *et al.* [62] place fictitious charges on each atom, which are then used to obtain dipole vectors. A similar approach is also discussed in chapter 6, published in [20], where response operators are directly incorporated into the learning process.

A representation should also contain as little superfluous information as possible. Such decluttering can be achieved by striving for a surjective mapping between compound and representation. If the mapping from compound to representation is both surjective and an injective,

then it is a bijection. A bijection means that there is precisely one compound for each representation. Following Ockham's razor, imposing the conditions mentioned above on a representation ensures that it uniquely represents a system while not containing any redundant information.

Smooth properties impose additional constraints on the structure of a representation. The mapping from a compound to a representation and its inverse should therefore also be smooth. Such a mapping is called a diffeomorphism, and a representation that fulfills this criterion is necessary to model physical quantities involving differentiation, such as forces.

Finding a representation which fulfills most, or all, of these criteria is non-trivial. However, as mentioned above, a suitable representation does not necessarily need to fulfill all criteria above if the study is restricted to a specific chemical compound subspace. For example, chapter 7 discusses a work, published in [8], with a limited chemical space. The chemical space in the study consists of a single crystal archetype where element values are substituted at each site. Therefore, it is sufficient to represent the system by a list containing the elemental identity of the species occupying each site.

## 2.6 Current State of the Field

Figure 6.4 contains LCs of several QML models published over the past decade on the QM9 [63] dataset. Because the sampling of the training data, test data, and hyper-parameterization might differ between works, the results should be compared with some caution. Despite this caveat, the figure provides an overview of how the field has advanced over the years, where significant year-by-year progress in model performance is evident.

As is now clear, the specific design of a QML model directly impacts its predictive accuracy. For example, passing a representation of each atom $A$ to an atomic QML model $f_{atom}(A, \alpha)$ which is then summed up $f(C, \alpha) = \sum_{A \in C} f_{atom}(A, \alpha)$, as opposed to passing a representation of the entire system $C$ to a QML model, is now common practice for extensive properties. This is because, in contrast to intensive properties, extensive properties grow with the system size (number of atoms).

Consequently, QML models formulated as a sum of atomic contributions have proven to yield remarkable accuracies for extensive properties such as atomization energies [19, 27, 38, 66, 72]. However, a QML model based upon a sum of atomic contributions would be a poor predictor of intensive properties, which do not scale with system size. Therefore, the pairing choice of a QML model architecture with the relevant target property heavily influences the model's

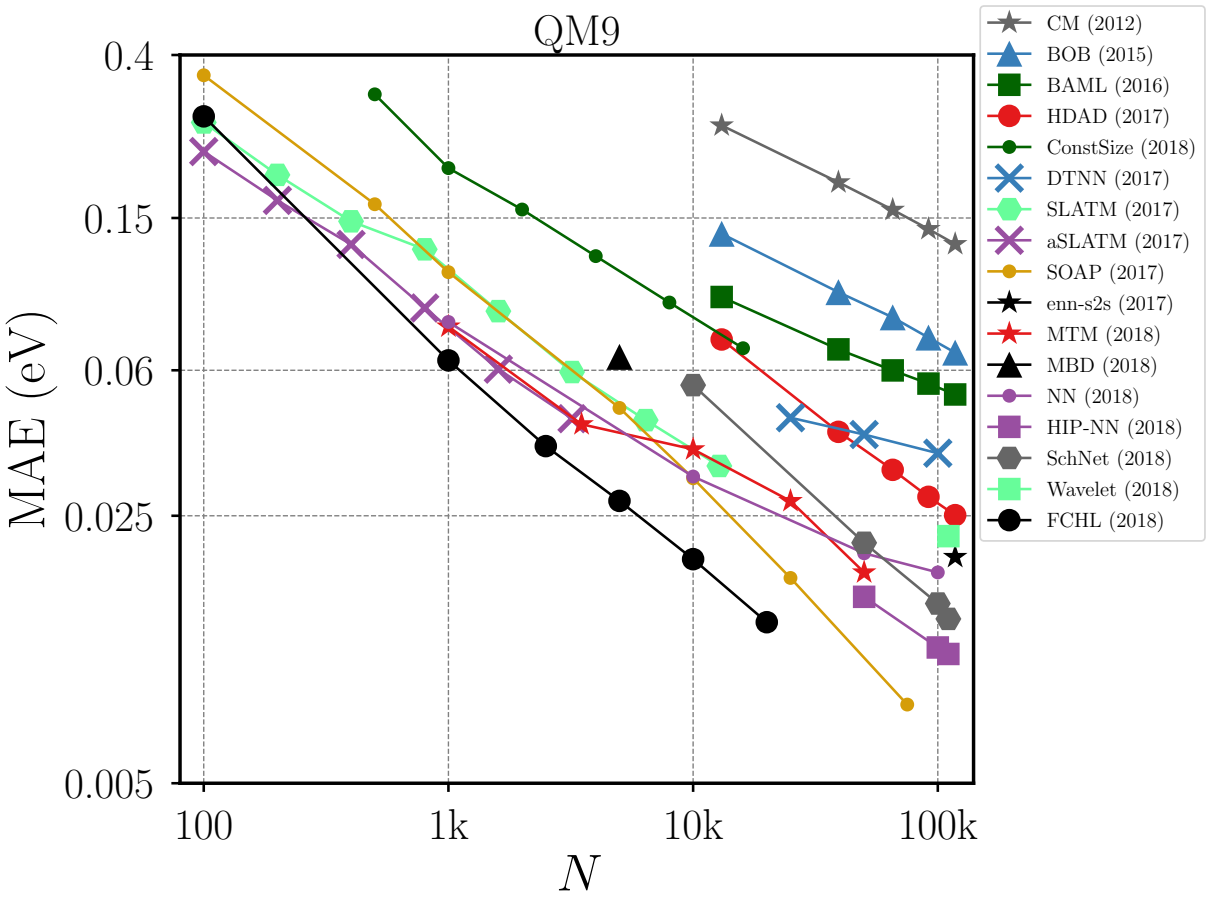



**Figure 6.4:** Performance overview of various QML models published ever since Ref. [33]. Prediction errors of atomization energies in the QM9 [63] (explained in chapter 3) are shown as a function of training set size. The QML models included differ solely by representation, model architecture and cross-validation details. The models correspond to CM [30, 33], BOB [64], BAML [56], HDAD [18], constant size [65], DTNN [41], (a)SLATM [66], SOAP [27], enn [39], MTM [67], MBD [68], NN [69], HIP-NN [70], SchNet [38], Wavelet [71], and FCHL [19].

predictive accuracy.

While extensive properties have been successfully predicted using QML models, the prediction of response properties and intensive properties has proven to be more challenging. That being said, tremendous effort has already been put into developing efficient QML models for response properties.

Perhaps the most relevant quantum response property in molecular dynamics simulations is the force exerted on an atom in a compound. Hence, remarkable strides have already been made, resulting in QML models which can predict *ab initio* quality forces at a low computational cost [28, 29, 38, 61, 73–81]. QML architectures with energy conserving force fields have also been proposed by several groups [60, 61], which is imperative in many statistical mechanics

applications.

Other response properties have also been investigated. The dipole moment, an important property in many applications, has been investigated thoroughly [41, 56, 62, 70, 82–85]. Schütt *et al.* [41] trained a neural network on the dipole vector itself, yielding a model with high accuracy on the QM9 dataset [63, 86]. NNs trained to predict atomic charges which are then used to estimate infrared intensities [62, 84] have also been investigated. Chapter 6, discusses a formalism for learning response properties of a given compound's energy with the help of response operators.

As a final note in this chapter, finding the ground-state electron density of a compound is also a highly coveted application of QML models, and several models have already been proposed. Brockherde *et al.* [87] use KRR to learn densities in a plane wave basis, and show that the densities can be used in molecular dynamics simulations. Sinitskiy and Pande [88] learns the electron densities of molecules taken from the QM9 dataset [63] with the help of a convolutional NN on a fixed grid. Grisafi *et al.* [89] presents a kernel model for efficient learning of electron densities in a local atomic basis.

# Chapter 3

# Chemical Space and Data-sets

In the broadest terms, the complete chemical space spans every combination of every atom type and configuration. Such a 'complete' chemical space spans infinite dimensions and consists of an endless set of compounds. Hence, this general chemical space therefore exists more as a philosophical understanding than as a useful tool; however, constricted subsets of the general space serve as helpful tools to analyze the behaviors of chemical reactions, substituent effects, conformer energies, or any other chemical feature that is best described over a spectrum of values. It is therefore only natural that this concept has proven useful in QML, were ML models are used to interpolate properties across various chemical subspaces, including composition and geometry. This thesis makes use of a relatively diverse set of structures to evaluate QML models, indicating how well a model performs across a given chemical subspace.

The data-sets used in this thesis include organic molecules, crystals, biomolecular dimers, water clusters, and main-group diatomics. All data-sets are either collected from literature or calculated in-house. The rest of this chapter outlines the most relevant data sets used in this thesis.

## 3.1 Organic molecules: QM9

The QM9 data-set [63] corresponds to the hybrid DFT [90] derived structures and properties of 134k organic drug-like molecules with up to nine heavy atoms (C, O, N, or F), not counting hydrogen.

Initial configurations correspond to SMILES strings, taken from a subset of the GDB-17 data-set [86]. Corina [91] was used to turn the SMILES strings to Cartesian coordinates. The geometries were then relaxed using PM7, followed by relaxation at B3LYP level of theory. Several properties were subsequently calculated for all molecules, including: energies and enthalpies

of atomization, HOMO and LUMO eigenvalues, Norm of dipole moment, static polarizability, zero-point vibrational energy (ZPVE), heat capacity at room temperature, and highest fundamental vibrational frequency.
QM9 is used in 4 and 5 to benchmark the performance of different machine learning models.

## 3.2 Organic molecules: QM7b

QM7b contains structures and properties of ∼7k organic drug-like molecules with up to seven heavy atoms (C, O, N, S or Cl), not counting H.
Similarly to QM9, the QM7b data-set [82] was derived from SMILES strings taken from the OpenBabel [92], which were then relaxed using hybrid DFT (PBE0 [93, 94]).
QM7b is used in chapter 5 to benchmark the predictive accuracy of QML models resulting from the representation introduced in the chapter.

## 3.3 Biomolecular dimers: SSI

A subset of 2356 neutral biomolecular dimers from the SSI data-set [95] is used in chapter 5 to benchmark the QML models for intra-molecular and non-equilibrium interactions.
The SSI data-set is a collection of dimers mimicking configurations of interacting amino-acid sidechains, obtained from a set of experimentally observed 47 high-resolution crystalline protein structures. The interaction energies were calculated using DW-CCSD(T**)-F12 level of theory [96].

## 3.4 Water cluster

A water cluster data-set is used in chapter 5 to evaluate the performance of QML models in a simulated water droplet. The data-set consists of 4000 configurations, each containing 40 water molecules.
The water cluster was generated by performing a molecular dynamics simulation of a 20 Åradius water shell at 300K with the standard stochastic boundary condition[97] in the CHARMM program[98] version c41a1. The structure of the water molecules was simulated using a modified TIP3P model[99, 100]. Non-bonded interactions were treated using extended electrostatics[101] and a switching function[102] for the van der Waals interaction between 8 and 12 Å. The duration of the simulation was 4ns with a 1fs time step. Snapshots were saved every 1000 time steps. In each snapshot, the coordinates of the 40 water molecules closest to the center of the

water cluster were saved, and the energy was calculated at the QM level using the PBEh-3c[103] method.

## 3.5 Solids: OQMD

The OQMD data-set is used to benchmark the the performance of QML models in 5. The data-set consists of crystals with calculated properties by Wolverton and co-workers [104, 105], is a subset of the Inorganic Crystal Structure Database (ICSD) [106, 107]. This data-set has already been used to develop and benchmark random forest- and kernel-based QML models (Voronoi) [108]. The data-set consists of ∼30k crystal structures and formation energies, calculated using high-throughput DFT with generalized gradient approximation (GGA+U).

## 3.6 Solids: Elpasolites

For training and evaluation, DFT formation energies for two data sets of Elpasolites were generated: one small, (III−VI), made up from only 12 elements, C, N, O, Al, Si, P, S, Ga, Ge, As, Sn, and Sb; and one large, (I−VIII), containing all main-group elements up to Bi. Since (III−VI) only comprise ∼12 k possible permutations, the complete list of formation energies was obtained.
(I−VIII) consists of 10 k structures, i.e. 0.5% of the total number of 2 M possible crystals. The (I−VIII) data set was generated through a random selection of Elpasolites while ensuring an unbiased composition.
The crystal structures were processed using the high-throughput toolkit [109]. DFT, as implemented in the Vienna *ab-initio* simulation package (VASP 5.2.2) with projector augmented wave pseudopotentials (PAWs) [110–112], was used to carry out the structural relaxation, and to obtain unit cell energies. The exchange-correlation effects were treated using the functional of Perdew, Burke, and Ernzerhof (PBE) [113]. First, a low-accuracy relaxation of the cell volume and internal degrees of freedom was made, followed by repeated restarting of VASP relaxation runs until the final energy difference was below 10 meV/atom. In all calculations, a Monkhorst-Pack [114] $k$-mesh of at least 3x3x3, and an energy cutoff of the plane-wave basis of 600 eV was used. The formation energies were obtained as the differences between the elpasolite unit cell energies per atom and the ground state energies of stoicheometrically equivlaent elemental solids, calculated with the same VASP settings.
Elemental crystal structures used as input for these calculations were taken from Ref. [115].

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | -3.396 | | | | | | | | | | | | | | | He -1.170 |
| Li | -1.905 | Be | -3.743 | B | -6.680 | C | -9.226 | N | -8.340 | O | -4.952 | F | -1.909 | Ne | -2.798 | |
| Na | -1.311 | Mg | -1.602 | Al | -3.761 | Si | -5.393 | P | -5.378 | S | -4.137 | Cl | -1.849 | Ar | -6.967 | |
| K | -1.027 | Ca | -1.981 | Ga | -3.027 | Ge | -4.618 | As | -4.657 | Se | -3.489 | Br | -1.634 | Kr | -5.399 | |
| Rb | -0.933 | Sr | -1.687 | In | -2.722 | Sn | -4.003 | Sb | -4.124 | Te | -3.144 | I | -1.517 | Xe | -3.583 | |
| Cs | -0.854 | Ba | -1.923 | Ti | -2.364 | Pb | -3.741 | Bi | -4.039 | | | | | | | |

**Table 6.1:** Calculated total energies [eV/atom] of elemental crystals used to obtain formation energies. Input crystal structures were taken from Ref. [115].

Results are shown in table 6.1.

## 3.7 MD snapshots: MD17 and ISO17

MD17 and ISO17 are data-sets consisting of snapshots of MD trajectories and is used in chapter 6 to compare the operator formalism to other machine learning models.

The MD17 is built up of 150k to almost 1M geometries of different molecules, sampled of MD trajectories [61]. The molecules include Benzene, Uracil, Naphthalene, Asprin, Salicylic acid, Malonaldehyde, Ethanol, and Toluene. The MD trajectories run at 500 Kelvin and at a time-resolution of 0.5 fs and are calculated at PBE+vdW-TS [116, 117] level of theory.

ISO17 [38, 118] consists of MD snapshots of 113 isomers, randomly taken from the $C_7O_2H_{10}$ isomer subset of QM9 [63]. The MD trajectories run at 500 Kelvin and at a time-resolution of 1 fs and are calculated using PBE [116] level of theory. The ISO17 comes with two validation sets. The first consists only of isomers with connectivity present in the training set. The second consists only of isomers with connectivity absent from the training set.

# Chapter 4

# Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error



## 4.1 Executive Summary

We investigate the impact of choosing regressors and molecular representations for the construction of fast machine learning (ML) models of thirteen electronic ground-state properties of organic molecules. The performance of each regressor/representation/property combination is assessed using LCs which report out-of-sample errors as a function of training set size with up to ∼118k distinct molecules. Molecular structures and properties at hybrid density functional theory (DFT) level of theory come from the QM9 database [Ramakrishnan et al, *Scientific Data* **1** 140022 (2014)] and include enthalpies and free energies of atomization , HOMO/LUMO energies and gap, dipole moment, polarizability, zero point vibrational energy, heat capacity and the highest fundamental vibrational frequency. Various molecular representations have been studied (Coulomb matrix, bag of bonds, BAML and ECFP4, molecular graphs (MG)), as well as newly developed distribution based variants including histograms of distances (HD), and angles (HDA/MARAD), and dihedrals (HDAD). Regressors include linear models (Bayesian ridge regression (BR) and linear regression with elastic net regularization (EN)), random forest (RF), kernel ridge regression (KRR) and two types of neural networks, graph convolutions (GC) and gated graph networks (GG). Out-of sample errors are strongly dependent on the choice of

representation *and* regressor *and* molecular property. Electronic properties are typically best accounted for by MG and GC, while energetic properties are better described by HDAD and KRR. The specific combinations with the lowest out-of-sample errors in the ∼118k training set size limit are (free) energies and enthalpies of atomization (HDAD/KRR), HOMO/LUMO eigenvalue and gap (MG/GC), dipole moment (MG/GC), static polarizability (MG/GG), zero point vibrational energy (HDAD/KRR), heat capacity at room temperature (HDAD/KRR), and highest fundamental vibrational frequency (BAML/RF). We present numerical evidence that ML model predictions deviate from DFT (B3LYP) less than DFT (B3LYP) deviates from experiment for all properties. Furthermore, out-of-sample prediction errors with respect to hybrid DFT reference are on par with, or close to, chemical accuracy. The results suggest that ML models could be more accurate than hybrid DFT if explicitly electron correlated quantum (or experimental) data was available.

This work was done in collaboration with Google, which was responsible for running most of the calculations, optimizing the hyperparameters, and designing the neural network architectures. I was mainly responsible for analysing and drawing conclusions from the results. I also assisted in generating several of the representations and wrote the main body of the manuscript.

## 4.2 Introduction

Due to its favorable trade-off between accuracy and computational cost, Density Functional Theory (DFT) [9, 10] is the workhorse of quantum chemistry [119]—despite its well known shortcomings regarding spin-states, van der Waals interactions, and chemical reactions [120, 121]. Failures to predict reaction profiles are particularly worrisome [122], and recent analysis casts even more doubts on the usefulness of DFT functionals obtained through parameter fitting [123]. The prospect of universal and computationally much more efficient ML models, trained on data from experiments or generated at higher levels of electronic structure theory such as post-Hartree Fock or quantum Monte Carlo (e.g. exemplified in Ref. [31]), therefore represents an appealing alternative strategy. Not surprisingly, a lot of recent effort has been devoted to developing ever more accurate ML models of properties of molecular and condensed phase systems.

Several ML studies have already been published using a data set called QM9 [63], consisting of molecular quantum properties for the ∼134k smallest organic molecules containing up to 9 heavy atoms (C, O, N, or F; not counting H) in the GDB-17 universe [86]. Some of these studies have developed or used representations we consider in this work, such as BAML (Bonds, angles, ML) [56], bag of bonds (BOB) [64, 83] and the Coulomb matrix (CM) [33, 83]. Atomic variants of the CM have also been proposed and tested on QM9 [124]. Other representations have also been benchmarked on QM9 (or QM7 which is a smaller but similar data set), such as Fourier series of radial distance distributions [125], motifs [58], the smooth overlap of atomic positions (SOAP) [29] in combination with regularized entropy match [26], constant size descriptors based on connectivity and encoded distance distributions [126]. Ramakrishnan *et al.* [31] introduced a Δ-ML approach, where the difference between properties calculated at coarse/accurate quantum level of theories is being modeled. Furthermore, neural network models, as well as deep tensor neural networks, have recently been proposed and tested on the same or similar data sets [41, 81]. Dral *et al.* [127] use such data to machine learn optimal molecule specific parameters for the OM2 [128] semiempirical method, and orthogonalization tests are benchmarked in Ref. [129].

However, limited work has yet been done in systematically assessing *various* methods *and* properties on large sets of the exact same chemicals [30]. In order to unequivocally establish if ML has the potential to replace hybrid DFT for the screening of properties, one has to demonstrate that ML test errors are systematically lower than estimated hybrid DFT accuracies for all the

properties available. This study accomplishes that through a large scale assessment of unprecedented scale: (i) In order to approximate large training set sizes $N$, we included 13 quantum properties from up to ∼118k molecules in training (90% of QM9). (ii) We tested multiple regressors (Bayesian ridge regression (BR), linear regression with elastic net regularization (EN), random forest (RF), kernel ridge regression (KRR), neural network (NN) models graph convolutions (GC) [130] and gated graphs (GG) [131]) and (iii) multiple representations including BAML, BOB, CM, extended connectivity fingerprints (ECFP4), histograms of distance, angle, and dihedral (HDAD), molecular atomic radial angular distribution (MARAD), and molecular graphs (MG). (iv) We investigated *all* combinations of regressors and representations, except for MG/NN which was exclusively used together because GC and GG depend fundamentally on the input representation being a graph instead of a flat feature vector.

The best models for the various properties are: atomization energy at 0 Kelvin (HDAD/KRR), atomization energy at room temperature (HDAD/KRR), enthalpy of atomization at room temperature (HDAD/KRR), atomization of free energy at room temperature (HDAD/KRR), HOMO/LUMO eigenvalue and gap (MG/GC), dipole moment (MG/GC), static polarizability (MG/GG), zero point vibrational energy (HDAD/KRR), heat capacity at room temperature (HDAD/KRR), and the highest fundamental vibrational frequency (BAML/RF). For training set size of ∼118k (90% of data set) we have found the additional out-of-sample error added by ML to be lower or as good as DFT errors at B3LYP level of theory relative to experiment for all properties, and that chemical accuracy (See table 4.4) is reached, or in sight.

This paper is organized as follows: First we will briefly describe the methods, including data set, model validation protocols, representations, and regressors. In section III, we present the results and discuss them, and section IV concludes the paper.

## 4.3 Method

### 4.3.1 Data set

We have used the QM9 data set consisting of ∼134k drug-like organic molecules [63]. Molecules in the data set consist of H, C, O, N and F, and contain up to 9 heavy atoms. For each molecule several properties, calculated at DFT level of theory (B3LYP/6-31G(2df,p)), were included. We used: Atomization energy at 0 Kelvin $U_0$ (eV); atomization energy at room temperature $U$ (eV); enthalpy of atomization at room temperature $H$ (eV); atomization of free energy at room temperature $G$ (eV); HOMO eigenvalue $\varepsilon_{\mathrm{HOMO}}$ (eV); LUMO eigenvalue

$\varepsilon_{\mathrm{LUMO}}$ (eV); HOMO-LUMO gap $\Delta\varepsilon$ (eV); norm of dipole moment $\mu = \sqrt{\sum_{r\in x,y,z}(\int d\mathbf{r} n(\mathbf{r}) r)^2}$ (Debye), where $n(\mathbf{r})$ is the molecular charge density distribution; static isotropic polarizability $\alpha = \frac{1}{3}\sum_{i\in x,y,z}\alpha_{ii}$ (Bohr$^3$), where $\alpha_{ii}$ is the diagonal element of the polarizability tensor; zero point vibrational energy ZPVE (eV); heat capacity at room temperature $C_{\mathrm{v}}$ (cal/mol/K); and the highest fundamental vibrational frequency $\omega_1$ (cm$^{-1}$). For energies of atomization ($U_0$, $U$, $H$ and $G$) all models yield very similar errors. We will therefore only discuss $U_0$ for the remainder. The 3053 molecules specified in Ref. [63] which failed SMILES consistency tests were excluded from our study, as well as two linear molecules, leaving ∼131k molecules.

### 4.3.2 Model validation

Starting from the ∼131k molecules in QM9 after removing the ∼3k molecules (see above) we have created a number of train-validation-test splits. We have splitted the data set into test and non-test sets and varied the percentage of data in test set to explore the effect of amount of data in error rates. Inside the non-test set, we have performed 10 fold cross validation for hyperparameter optimization. That is, for each model 90% (the training set) of the non-test set is used for training and 10% (the validation set) is used for hyperparameter selection. For each test/non-test split, we have trained 10 models on different subsets of the non-test set, and we report the mean error on the test set across those 10 models. Note that the non-test set will be referred to as training set in the results section in order to simplify discussion.
In terms of CPU investments necessary for training the respective models we note that EN/BR, RF/KRR, and GC/GG required minutes, hours, and multiple days, respectively. Using GPUs could dramatically reduce such timings.

### 4.3.3 DFT errors

To place the quality of our prediction errors in the right context, experimental accuracy estimates of hybrid DFT become desirable. Here, we summarize literature results comparing DFT *at B3LYP level of theory* to experiments for the various properties we study. Where data is available, the corresponding deviation from experiment is listed in Table 4.4, alongside our ML prediction errors (*vide infra*).
In order to also get an idea of hybrid DFT energy errors for organic molecules, such as the compounds studied herewithin, we refer to a comparison of PBE and B3LYP results for 6k constitutional isomers of $C_7H_{10}O_2$ [31]. After centering the data by subtracting their mean shift from G4MP2 (177.8 (PBE) and 95.3 (B3LYP) kcal/mol). The remaining MAEs are

roughly ∼2.5 and ∼ 3.0 kcal/mol for B3LYP and PBE, respectively. This is in agreement with what Curtiss *et al.* [132] found. They compared DFT to experimental values from 69 small organic molecules (of which 47 were substituted with F, Cl, and S), with up to 6 heavy atoms (not counting hydrogens), and calculated the energies using B3LYP/6-311+G(3df,2p). The resulting mean absolute deviation from experimental values was 2.3 kcal/mol.

Rough hybrid DFT error estimates for dipole moment and polarizability have been obtained from Refs. [133]. The errors are estimated referenced to experimental values, for a data set consisting of 49 molecules with up to 7 heavy atoms (C, Cl, F, H, O, P, or S)

Frontier molecular orbital energies (HOMO, LUMO and HOMO-LUMO gap) can not be measured directly.

However, for the exact (yet unknown) exchange-correlation potential, the Kohn-Sham HOMO eigenvalues correspond to the negative of the vertical ionization potential (IP) [134]. Unfortunately, within hybrid DFT, the precise meaning of the frontier eigenvalues and the gap is less clear, and we therefore refrain from a direct comparison of B3LYP to experimental numbers. Nevertheless, we have included eigenvalues and the gap due to their widespread use for molecular and materials design applications.

Hybrid DFT RMSE estimates with respect to experimental values of ZPVE and $\omega_1$ (the highest fundamental vibrational frequency) were published in Ref. [135] for a set of 41 organic molecules, with up to 6 heavy atoms (not counting hydrogen) and calculated using B3LYP/cc-pVTZ.

Normally distributed data has a constant ratio between RMSE and MAE, [136] which is roughly 0.8. We have used this ratio to approximate the MAE from the RMSE estimates reported for ZPVE and $\omega_1$.

Deviation of DFT (at the B3LYP/6-311g** level of theory) from experimental heat capacities were reported by DeTar [137] who obtained errors of 16 organic molecules, with up to 8 heavy atoms (not counting hydrogens).

Note, however, that one should be cautious when referring to these errors: Strictly speaking they can not be compared since different basis sets, molecules, and experiments were used. We also note that all DFT errors in this paper are estimated from B3LYP and using other functionals can yield very different errors.

Nevertheless, we feel that the quoted errors provide meaningful guidance as to what one can expect from DFT for each property.

### 4.3.4 Representations

The design of molecular representations is a long-standing problem in chem-informatics and materials informatics, and many interesting and promising variants have already been proposed. Below, we provide the details on the representations selected for this study. While finalizing our study, competitive alternatives were introduced [27, 32] but have been tested only for energies (and polarizabilities).

#### CM and BOB

The Coulomb matrix (CM) representation[33] is a square atom by atom matrix, where off diagonal elements are the Coulomb nuclear repulsion terms between atom pairs. The diagonal elements approximate the electronic potential energy of the free atoms. Atom indices in the CM are sorted by the $L^1$ norm of each atom's row (or column). The Bag of Bonds (BOB)[64] representation uses exclusively CM elements, grouping them for different atom pairs into different bags, and sorting them within each bag by their relative magnitude.

#### BAML

The recently introduced BAML (Bonds, angles, ML) representation can be viewed as a many-body extension of BOB[56]. All pairwise nuclear repulsions are replaced by Morse/Lennard-Jones potentials for bonded/non-bonded atoms respectively. Furthermore, three- and four-body interactions between covalently bonded atoms are included using angular and torsional terms, respectively. Parameters and functional forms are based on the universal force field (UFF)[138].

#### ECFP4

Extended Connectivity Fingerprints [57] (ECFP4) are a common representation of molecules in cheminformatics based studies. They are particularly popular for drug discovery [139–141]. The basic idea, typical also for other cheminformatics descriptors [142] (e.g. the *signature* descriptor [143, 144]) is to represent a molecule as the set of subgraphs up to a fixed diameter (here we use ECFP4, which is a max diameter of 4 bonds). To produce a fixed length vector, the subgraphs can be hashed such that every subgraph sets one bit in the fixed length vector to 1. In this work, we use a fixed length vector of size 1024. Note that ECFP4 is based solely on the molecular graph specifying all covalent bonds, e.g. as encoded by SMILES strings.

**MARAD**

Molecular atomic radial angular distribution (MARAD) is an RDF based representation. Per atom it consists of three RDFs using Gaussians of interatomic distances, and parallel and orthogonal projections of distances in atom triplets, respectively. Distances between two molecules can be evaluated analytically. Unfortunately, most regressors evaluated in this work, such as BR, EN and RF, do not rely on inner products and distances between representations. We resolve this issue by projecting MARAD onto bins in order to work with all regressors (apart for GG and GC which use MG exclusively). The three body terms in MARAD contain information about both, angles and distances of all atoms involved. This differs from HDA(see below), where distances, and angles are decoupled, and placed in separated bins. Note that unlike BAML or HDAD, there are only two and three-body terms, no four-body terms (dihedral angles) have been included within MARAD.
The environment of an atom $I$ is represented by three functions: $A_r(I)$ , $A_\perp(I)$ and $A_\parallel(I)$, see Eq. 3.1.

$$A_k(I) = \mathcal{Z}(R_I, \sigma_R; \chi_1)\mathcal{Z}(C_I, \sigma_C; \chi_2) \sum_i^{n_I} \Phi_i^k(I) \exp\left[-\frac{(\chi_3 - d_{i,I})^2}{2\sigma_d^2}\right] \mathcal{Z}(R_i, \sigma_R; \chi_4)\mathcal{Z}(C_i, \sigma_C; \chi_5)\xi(d_{i,I}) \tag{3.1}$$

$\chi$ is integrated out when comparing two atoms (or molecules), or when discretizing the representation; $\sigma_d$, $\sigma_R$, $\sigma_C$ are hyper parameters; $d_{i,I}$ is the distance between atom $I$ and nearby atoms i; $R_i$ and $C_i$ correspond respectively to the row and column of atom $i$ in the periodic table; $\xi(d_{i,I})$ is a scaling function that is used to give a higher importance to smaller distances, as chemical bonds in molecules are mostly affected by nearby atoms; and $\mathcal{Z}(R, \sigma; \chi)$ is used to introduce a chemical similarity between two atoms of different, or the same, elemental type. $\Phi_i^k(I)$, is equal to 1, $\sum_j \cos(\theta_{i,j}^I)\xi(d_{i,I})$ and $\sum_j \sin(\theta_{i,j}^I)\xi(d_{i,I})$ for $k = r$, $\parallel$ and $\perp$ respectively. $\theta_{i,j}(I)$ is the unsigned angle between the vector spanning from atom I to atom i and the vector spanning from atom I to atom j.
Most regressors used in this chapter, such as BR, EN and RF, do not rely on inner products and distances between representations. Therefore, we generated MARAD $M_k$ by summing $A_k(I)$ over all $n$ atoms $I$ in the molecules, which we discretize by calculating the scalar product between $M_k$ and a grid.
The grid points $\mathcal{G}_{i,a,b}$ are placed with uniform spacing $\sigma_d$ along the interatomic distances $d$, on the row $R$ and column $C$ in the periodic table for each element pair.

$$\mathcal{S}(M_k, \mathcal{G}_{j,a,b}) \equiv \sum_I^n \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} A_k(I)\mathcal{G}_{j,a,b}^{proj} d\chi_1 \cdots d\chi_5 \tag{3.2}$$

$$\mathcal{G}_{j,a,b} = \mathcal{Z}(R_a, \sigma_R; \chi_1)\mathcal{Z}(C_a, \sigma_C; \chi_2)\exp(-\frac{(\chi_3 - \sigma_d j)^2}{2\sigma_d^2})\mathcal{Z}(R_b, \sigma_R; \chi_4)\mathcal{Z}(C_b, \sigma_C; \chi_5)\xi(\sigma_d j) \tag{3.3}$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} A_k(I)\mathcal{G}^{proj}_{i,a,b}d\chi_1\cdots d\chi_5 = \frac{\sqrt{\pi}\sigma_d\sigma_R^4\sigma_C^4}{(\sigma_R^2 + (R_I - R_a)^2)(\sigma_C^2 + (C_I - C_a)^2)} \sum_i^{n_I}\Phi_i^k(I)\frac{\exp(-\frac{(\sigma_d j - d_{i,I})^2}{4\sigma_d^2})\xi(d_{i,I})\xi(\sigma_d j)}{(\sigma_R^2 + (R_i - R_b)^2)(\sigma_C^2 + (C_i - C_b)^2)} \tag{3.4}$$

Throughout this chapter, the hyperparameters $\sigma_R$, $\sigma_C$ and $\sigma_d$ were set equal to 1, 0.5 and 0.2 respectively, and a sinusoidal scaling function with a hard cutoff was used: $\xi(d) = 1 - \sin(\pi\frac{d}{2D})$ if $d \leq D$ and 0 otherwise, with a cutoff distance $D = 6$ Å. The chemical similarity was set equal to $\frac{\sigma^{3/2}}{\sqrt{\pi}2[(\sigma/2)^2 + (\chi - R)^2]}$

**HD, HDA, and HDAD**

BOB, BAML and MARAD rely on computing functions for given interatomic distances, and/or angles, and/or torsions, and then either project that value on to discrete bins, or sort the values. As a straightforward alternative, we also investigated representations which account directly from pairwise distances, triple-wise angles, and quad-wise dihedral angles through manually generated bins in histograms. The resulting representations in increasing interatomic many-body order are called HD(Histogram of distances), HDA (Histogram of distances and angles), and HDAD (Histogram of distances, angles and dihedral angles). For any given molecule, one iterates through each atom $a_i$, producing a set of distances, angle and dihedral angle features for $a_i$.

*Distance features* were produced by measuring the distance between $a_i$ and $a_j$ (for $i \neq j$) for each element pair. The distance features were assigned a label incorporating the atomic symbols of $a_i$ and $a_j$ sorted alphabetically (with H last), e.g. if $a_i$ was a carbon atom and $a_j$ was a nitrogen atom, the distance feature for the atom pair would be labeled C-N. These labels will be used to group all features with the same label into a histogram and allow us to only count each pair of atoms once.

*Angle features* were produced by taking the principal angles formed by the two vectors spanning from each atom $a_i$ to every subset of 2 of its 3 nearest atoms, $a_j$ and $a_k$. The angle features were labeled by the element type of $a_i$, followed by the alphabetically sorted element types (Except for hydrogens, which were listed last) of $a_j$ and $a_k$. The example where $a_i$ is a Carbon atom, $a_j$ a Hydrogen atom, $a_k$ a Nitrogen would be assigned the label C-N-H.

*Dihedral angle features* were produced by taking the principal angles between two planes. We take $a_i$ as the origin, and for each of the four nearest neighbors in turn, labeling the neighbor atom $a_j$, and forming a vector $V_{ij} = a_i \rightarrow a_j$. Then all $\binom{3}{2}$ subsets of the remaining three out of four nearest neighbors of $a_i$ are chosen, and labeled as $a_k$ and $a_l$. This third and fourth atom respectively form two triangular faces when paired with $V_{ij}$: $\langle a_k, a_i, a_j \rangle$ and $\langle a_l, a_i, a_j \rangle$. The dihedral angle between the two triangular faces was calculated. These dihedral angle features were labeled with the atomic symbol for $a_i$, followed by the atomic symbols for $a_j$, $a_k$ and $a_l$, sorted alphabetically, with the exception that hydrogens were listed last, e.g. C-C-N-H.

The features from all molecules have been aggregated for each label type to generate a histograms for each label type. Fig. 3.1 exemplifies this for C-N distances, C-C-C angles, and C-C-C-O dihedrals for the entire QM9 data set. Certain typical molecular characteristics can be recognized upon mere inspection. For example, the CN histogram displays a strong and isolated peak between 1.1 and 1.5 Å, corresponding to occurrences of single, double, and triple bonds. For distances above 2 Å, peaks at typical radii of second and third covalent bonding shells around N can be recognized at 2.6 Å and 3.9 Å. Also C-C-C angles can be easily interpreted: The peak close to zero and $\pi$ Rad corresponds to geometries where three atoms are part of a linear (alkyne, or nitrile) type of motif. The broad and largest peak corresponds to 120 and 109 degrees, typically observed in $sp^2$ and $sp^3$ hybridized atoms.

The morphology of each histogram has then been examined to identify apparent peaks and troughs, motivated by the idea that peaks indicate structural commonalities among molecules. Bin centers have been placed at each significant local minimum and maximum (Shown as vertical lines in Fig. 3.1). Values at 15-25 bin centers have been chosen as a representation for each label type. For each molecule, the collection of features has subsequently been rendered into a fixed-size representation, producing one vector component for each bin center, within each label type. This has been accomplished following a two-step process. (i) *Binning and interpolation:* Each feature value is projected on the two nearest bins. The relative amount projected on each bin uses linear projection between the two bins. For example: A feature with value 1.7 which lies between two bins placed at 1.0 and 2.0 respectively, contributes 0.3 and 0.7 to the first and second bin respectively. (ii) *Reduction:* The collection of contributions within each bin of each molecule's feature vector is condensed to a single value by summing all contributions.

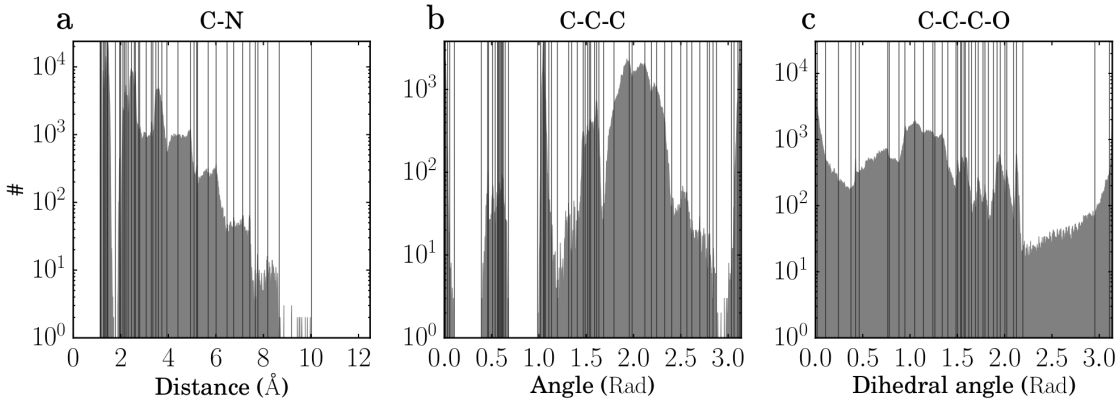



**Figure 3.1:** Illustration of select histograms of distances, angles and dihedral angles in QM9. Vertical lines constitutes placements of the bins in the HDand/or HDAD representations. (a) All C-N distances. (b) All C-C-C angles. (c) All C-C-C-O dihedral angles.

### Molecular Graphs

We have investigated several neural network models which are based on molecular graphs (MG) as representation. The inputs are real-valued vectors associated with each atom and with each pair of atoms. More specifically, we have used the featurization described in Kearnes *et al.* [130] with the removal of partial charge and the addition of Euclidean distances to the pair feature vectors. All elements of the feature vector are described in Tables 3.1 and 3.2.

The featurization process was unsuccessful for a small number of molecules (367) because of conversion failures from geometry to rational SMILES string when using OpenBabel [145] or RDKit [146], and were excluded from all results using the molecule graph features.

**Table 3.1:** Atom features for the MG representation: Values provided for each atom in the molecule.

| Feature | Description |
|---|---|
| Atom type | H, C, N, O, F (one-hot). |
| Chirality | R or S (one-hot or null). |
| Formal charge | Integer electronic charge. |
| Ring sizes | For each ring size (3–8), the number of rings that include this atom. |
| Hybridization | *sp*, $sp^2$, or $sp^3$ (one-hot or null). |
| Hydrogen bonding | Whether this atom is a hydrogen bond donor and/or acceptor (binary values). |
| Aromaticity | Whether this atom is part of an aromatic system. |

Note that within a previous draft of this study [147], we reported biased results for GC/GG models due to use of Mulliken partial charges within the MG representation. All MG results

**Table 3.2:** Atom pair features for the MG representation: Values provided for each pair of atoms in the molecule.

| Feature | Description |
|---|---|
| Bond type | Single, double, triple, or aromatic (one-hot or null). |
| Graph distance | For each distance (1–7), whether the shortest path between the atoms in the pair is less than or equal to that number of bonds (binary values). |
| Same ring | Whether the atoms in the pair are in the same ring. |
| Spatial distance | The euclidean distance between the two atoms. |

presented herewithin have been obtained without any Mulliken charges in the representation. Model hyper parameters for the GC model, however, still correspond to the previously obtained hyper parameter search.

### 4.3.5 Regressors

For all methods, we first standardized the property values so that all properties have zero mean and unit standard deviation.

#### Kernel Ridge Regression

KRR [22–25] is a type of regression with regularization [148] which uses kernel functions as basis set. A property $p$ of a query molecule $\mathbf{m}$ is predicted by a sum of weighted kernel functions $K(\mathbf{m}, \mathbf{m}_i^{\text{train}})$ between $\mathbf{m}$ and all molecules $\mathbf{m}_i^{\text{train}}$ in the training set,

$$p(\mathbf{m}) \quad = \quad \sum_{i}^{N} \alpha_i K(\mathbf{m}, \mathbf{m}_i^{\text{train}}) \tag{3.5}$$

where $\alpha_i$ are regression coefficients, obtained by minimizing the Euclidean distance between the estimated and the reference property of all molecules in the training set. We used Laplacian and Gaussian kernels as implemented by scikit-learn [149] for all representations.
The level of noise in our data is very low so strong regularization is not necessary. We have set the regularization parameter to $10^{-9}$, and we note that prediction errors change negligibly when altering it to $10^{-10}$. Kernel widths were chosen by screening values on a base-2 logarithmic grid for the 10 percent training set (from 0.25 to 8192 for Gaussian kernel and 0.1 to 16384 for Laplacian kernel). In order to simplify the width screening, prior to learning all feature vectors

were normalized (scaling the input vector by the mean norm across the training set) by the Euclidean norm for the Gaussian kernel and the Manhattan norm for the Laplacian kernel.

**Bayesian Ridge Regression**

We use BR [150] as is implemented in scikit-learn [149]. BR is a linear model with a $L^2$ penalty on the coefficients. Unlike Ridge Regression where the strength of that penalty is a regularization hyperparameter which must be set, in Bayesian Ridge Regression the optimal regularizer is estimated from the data.

**Elastic Net**

Also EN [151] is a linear model. Unlike BR, the penalty on the weights is a mix of $L^1$ and $L^2$ terms. In addition to the regularization hyperparameter for the weight penalty, Elastic net has an additional hyperparameter l1_ratio to control the relative strength of the $L^1$ and $L^2$ weight penalties. We used the scikit-learn [149] implementation and set l1_ratio $= 0.5$. We then did a hyperparameter search on regularizing parameter in a base 10 logarithmic grid from $1e-6$ to 1.0.

**Random Forest**

We use RF [152] as implemented in scikit-learn [149]. RF regressors produce a value by averaging many individual decision trees fitted on randomly resampled sets of the training data. Each node in each decision tree is a threshold of one input feature. Early experiments did not reveal strong differences in performance based on the number of trees used, once a minimal number was reached. We have used 120 trees for all regressions.

**Graph Convolutions**

We have used the GC model as described in Kearnes *et al.* [130], with several structural modifications and optimized hyperparameters. The graph convolution model is built on the concepts of "atom" layers (one real vector associated with each atom) and "pair" layers (one real vector associated with each pair of atoms). The graph convolution architecture defines operations to transform atom and pair layers to new atom and pair layers. There are three structural changes to the model used herewithin when compared to the one described in Kearnes *et al.* [130]. First, we have removed the "Pair order invariance" property by simplifying the $(A \to P)$ transformation. Since the model only uses the atom layer for the molecule level features, pair order invariance is not needed. Second, we have used the Euclidean distance between atoms. In

the $(P \to A)$ transformation, we divide the value from the convolution step by a series of distance exponentials. If the original convolution for an atom pair $(a, b)$ with distance $d$ produces a vector $V$, we concatenate the vectors $V$, $\frac{V}{d^1}$, $\frac{V}{d^2}$, $\frac{V}{d^3}$, and $\frac{V}{d^6}$ to produce the transformed value for the pair $(a, b)$. Third, we have followed other work on neural networks based on chemical graphs [40] which uses a sum of softmax operations to convert a real valued vector to a sparse vector and sum those sparse vectors across all the atoms. We use the same operation here along with a simple sum across the atoms to produce molecule level features from the top atom layer. We have found that this works as well or better than the Gaussian histograms first used in GC [130]. To optimize the network, we have searched the hyperparameter space using Gaussian Process Bandit Optimization [153] as implemented by HyperTune [154]. The hyperparameter search has been based on the evaluation of the validation set for a single fold of the data.

#### Gated Graph Neural Networks

We have used the GG Neural Networks model (GG) as described in Li *et al.* [131]. Similar to the GC model, it is a deep neural network whose input is a set of node features $\{x_v, v \in G\}$, and an adjacency matrix $A$ with entries in a discrete set $S = \{0, 1, \cdots, k\}$ to indicate different edge types. It has internal hidden representations for each node in the graph $h_v^t$ of dimension $d$ which it updates for $T$ steps of computation. Its output is invariant to all graph isomorphisms, meaning the order of the nodes presented to the model does not matter. To include the most relevant distance information we distinguish four different covalent bonding types (single, double, triple, aromatic). For all remaining atom-pairs we bin them by their interatomic distance [in Å] into 10 bins: [0, 2], [2,2.5], [2.5,3], [3,3.5], [3.5,4], [4,4.5], [4.5,5], [5,5.5], [5.5,6], and [6,$\infty$]. Using these bins, the adjacency matrix has entries in an alphabet of size 14 (k=14), indicating bond type for covalently bonded atoms, and distance bin for all other atoms. We have trained the GG model on each target property individually.

## 4.4 Results and discussion

### 4.4.1 Overview

We present an overview of the most relevant numerical results in Table 4.4. It contains the test errors for all combinations of regressors and representations and properties for models trained on $\sim$118 k molecules. The best models for the respective properties are $U_0$ (HDAD/KRR), $\varepsilon_{\text{HOMO}}$ (MG/GC), $\varepsilon_{\text{LUMO}}$ (MG/GC), $\Delta\varepsilon$ (MG/GC), $\mu$ (MG/GC), $\alpha$ (MG/GG), ZPVE (HDAD/KRR),

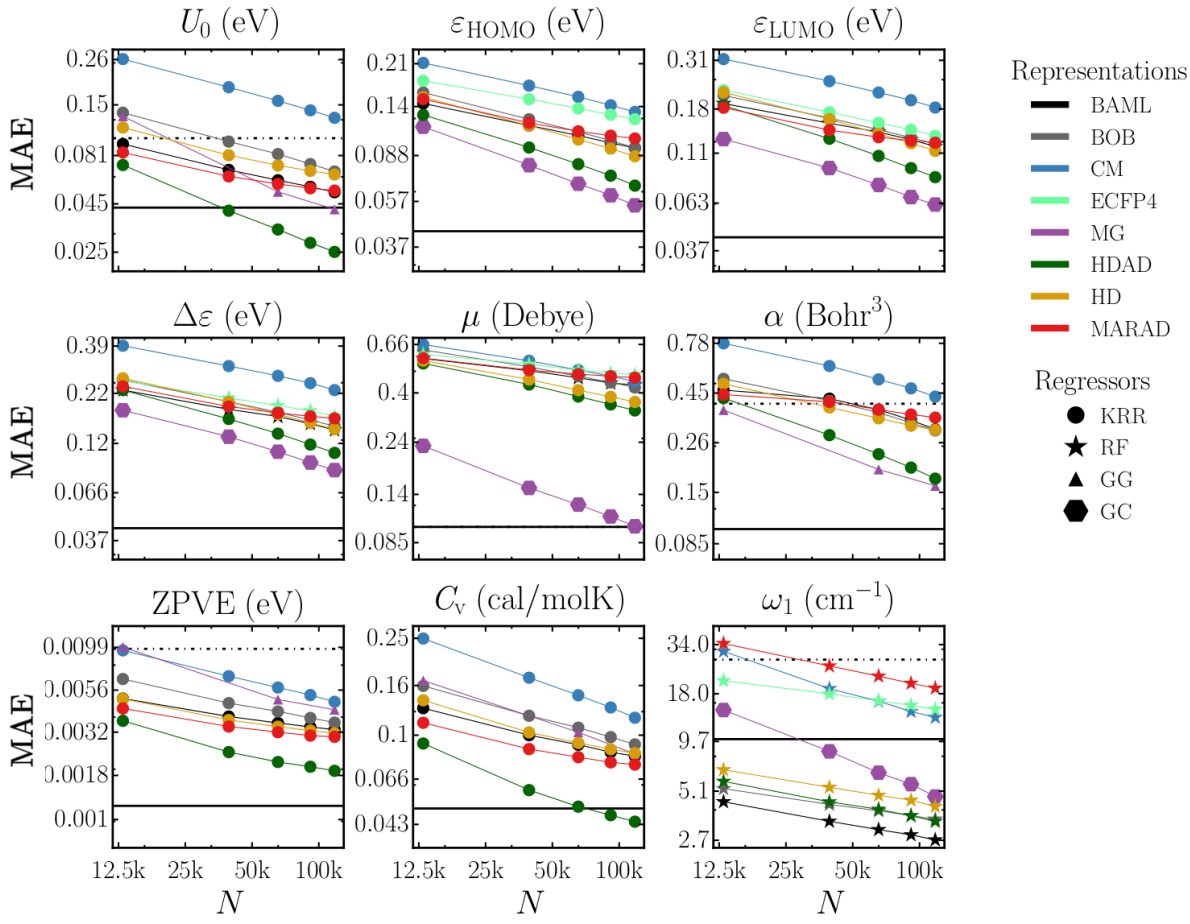



**Figure 4.2:** LCs (MAE as a function of training set size $N$) for 10 properties of QM9 molecules, displaying the best regressor for each representation and property. Horizontal solid lines correspond to target accuracies, vertical dotted lines correspond to approximated B3LYP accuracies (unless off-chart), see also table 4.4. Note that due to its poor performance ECFP4 results have been excluded for $\alpha$, ZPVE, U and $C_{\mathrm{v}}$.

$C_{\mathrm{v}}$ (HDAD/KRR), and $\omega_1$ (BAML/RF). We do not show results for the other three energies, $U(T = 298K), H(T = 298K), G(T = 298K)$ since identical observations as for $U_0$ can be made. Overall, NN and KRR regressors perform well for most properties. The ML out-of-sample errors outperform DFT accuracy at B3LYP level of theory and reach chemical (target) accuracy, both

defined alongside in table 4.4, for $U_0$ (HDAD/KRR and MG/GG), $\mu$ (GC), $C_\mathrm{v}$ (HDAD/KRR), and $\omega_1$ (BAML/KRR, MG/GC, HDAD/KRR, BOB/KRR, HD/KRR and MG/GG). For the remaining properties ( $\varepsilon_\mathrm{HOMO}$, $\varepsilon_\mathrm{LUMO}$, $\Delta\epsilon$, $\alpha$, and ZPVE) the best models come within a factor 2 of target accuracy, while all (except $\varepsilon_\mathrm{HOMO}$, $\varepsilon_\mathrm{LUMO}$ and $\Delta\epsilon$) where we don't have reliable data. outperforming DFT accuracy.

In Fig. 4.2 out-of-sample errors as a function of training set size (LCs) are shown for all properties and representations with the best corresponding regressor. It is important to note that *all* models on display systematically improve with training set size, exhibiting the typical linearly decaying behavior on a log-log plot [52, 56]. Errors for most models shown decay with roughly the same slopes, indicating similar exponents in the power-law of error decay. Notable exceptions, i.e. property models with considerably steeper LCs (Slopes and off-sets of all LCs can be found in Tables 4.6 and 4.7), are MG/GC for $\mu$, MG/GG and HDAD/KRR for $\alpha$, CM/KRR and BOB/KRR for $\langle \mathrm{R}^2 \rangle$, HDAD/KRR and MG/GG for $U_0$, and MG/GG for $\omega_1$. These results suggest that the specified representations capture particularly well the effective dimensionality of the corresponding property in chemical space.

### 4.4.2 Regressors

Inspection of Table 4.4 indicates that the regressors can roughly be ordered by performance, independent of property and representation: GC > GG > KRR > RF > BR > EN. It is noteworthy how EN, BR, and RF regressors perform substantially worse than GC/GG/KRR. The bad performance of EN and BR is due to their low model capacities. This can also be seen from the LCs of all regressors presented in Figure 4.3. The performance of BR and EN improves only slightly with increased training set size and even gets worse for some property/representation combinations. These two regressors also exhibit very similar LCs and BR performs only slightly better than EN for most combinations. The only clear exception to this rule is for ZPVE and $U_0$ together with HDAD, where BR performs significantly better than EN. Also, BR and EN errors rapidly converge to a constant w.r.t. training set size for all representations and properties, except for HDAD, which is the only representation which has a noteworthy improvement with increased training set size for some properties. The constant learning rates are not surprising as (a) the number of free regression parameters in BR and EN is relatively small and does not grow with training set size, and as (b) the underlying model is a linear combination with small flexibility. This behavior implies error convergence already for relatively small training sets.

RF performs poorly compared to GC, GG and KRR for all properties except for $\omega_1$, the highest

lying fundamental vibrational frequency in each molecule. For this property RF yields an astounding performance with out-of-sample errors as small as single digit $cm^{-1}$. B3LYP achieves a mean absolute error of only 28 $cm^{-1}$ with respect to experiment [135]. The distribution of $\omega_1$, Fig. 1 of reference [83], suggests a simple reason for this: There are three distinct peaks which correspond to typical C-H, N-H and O-H stretch vibrations in increasing order. Therefore the principal learning task in this property is to detect if there is an OH group, and if not if there is an NH group. If neither group is present, CH will yield the vibration with the highest frequency. As such, this is essentially about classifying which bonds are present in the molecule. RF works by fitting a decision tree to the target property. Each branch in the tree is based on an inequality of *one* entry in the representation. RF should therefore be able to identify which bonds are present in a molecule, simply by looking at the entries in the each element pair, and/or triplet bin of the representations. For RF, a fractional importance can be assigned to each input feature (the sum of all importances is 1.0). Analyzing the importance of the bins in HDAD of the RF model reveals that the three bins with highest importance are: O-H placed at 0.961 Å, N-H placed at at 1.01 Å and C-C-H at 3.138 radians with an importance of 0.587, 0.305 and 0.064 respectively. These three first bins constitute $\sim$96% of the prediction of $\omega_1$ and distances of the O-H and N-H bins are very similar to O-H and N-H bond lengths. C-C-H is placed on $\sim \pi$ radians which means that it has to correspond to a linear C-C-H (alkyne) chain which implies that the two carbons must be bonded by a triple bond (typically the C-H with the lowest $pK_a$ and the highest C-H stretch vibration). KRR performs remarkably well on average. For extensive energetic properties it yields the lowest overall errors in combination with HDAD and BOB, respectively. Its outstanding performance is not unsurprising in light of the multiple previous ML studies dealing with compositional as well as configurational spaces. The neural network flavors GC and GG, however, yield better performance on average, and the lowest errors for all electronic (mostly intensive) properties, i.e. dipole moment, polarizability, HOMO/LUMO eigenvalues and gaps. A possible explanation for this property dependent difference in performance between KRR and NN could be the inherent respective additive and multiplicative nature of these regressors. The energy being extensive, it is consistent with this picture that effective, quasi-particle based linear KRR based estimates have recently been reported to deliver very accurate predictions which can scale [66].

### 4.4.3 Representations

As one would expect, HDAD contains more relevant information and thus it always outperforms HD when using KRR. Tests also showed that an HDArepresentation systematically yields errors in between HDAD and HD, and similar observations hold for BR and EN regressor. In the case of RF, however, we observe little difference between HDAD and HD, and HD can even yield slightly lower errors than HDAD. In our opinion, this is due to the additional bins of angles and dihedrals rather adding noise than signal. By contrast, the separation of distances, angles and dihedral angles into different bins is not a problem for the KRR methods because the kernels used are purely distance based. This makes it possible for KRR to exploit the extra three- and four-body information in HDAD and to gain an advantage over HD. We note however that the remarkable performance of HDAD is possible despite its striking simplicity. As illustrated in Fig. 3.1 and discussed above, characteristic chemical behavior can be directly obtained by human inspection of HDAD. As such, HDAD corresponds to a representation very much "Occam's razor style". Unfortunately, due to its discrete nature and its origin in sorting distances, HDAD will suffer from lack of differentiability, which might limit its applicability when modeling forces or other non-equilibrium properties.

MARAD, containing similar information as HDA, performs similarly to BAML—yet, MARAD requires no prior knowledge about the physics encoded in the universal force-field such as electronic hybridization states, bond-order, or functional potential shapes (Morse, Lennard-Jones, harmonic angular potentials, or sinusoidal dihedrals). BOB and CM, previously state of the art, result in relatively poor performance.

ECFP4 produces out-of-sample errors on par or slightly better than CM/KRR for intensive properties ($\mu$, HOMO/LUMO eigenvalues and gap), however the model produces errors that are off-the-chart for all extensive properties ($\alpha$, ZPVE, $U_0$ and $C_{\mathrm{V}}$).

## 4.5 Conclusions

We have benchmarked many combinations of regressors and representations on the same QM9 data set consisting of ∼131k organic molecules. For all properties, the best ML model prediction errors reach the accuracy of DFT at B3LYP level with respect to experiment. For 7 out of 12 distinct properties (atomization energies, heat-capacity, $\omega_1$,

$\mu$) out-of-sample errors reach levels on par with chemical accuracy, or better, using a training set size of ∼118k (90% of QM9 data set) molecules. For the remaining properties $\alpha$, $\varepsilon_{\mathrm{HOMO}}$,

$\varepsilon_{\mathrm{LUMO}}$, $\Delta\epsilon$, and ZPVE, errors of the best models come within a factor 2 of chemical accuracy. Regressors EN, BR, and RF lead to rather high out-of-sample errors, while KRR and graph based NN regressors compete for the lowest errors. We have found that GC, GG, and KRR have best performance across *all* properties, except for the highest vibrational frequency for which RF performs best. There is no single representation and regressor combination that works best for all properties (though forthcoming work with further improvements to the GG based models indicates best in class performance across all properties [39]). For intensive electronic properties ($\mu$, HOMO/LUMO eigenvalues and gap) we have found MG/NN to yield the highest predictive power, while HDAD/KRR corresponds to the most accurate model for extensive properties ($\alpha$, ZPVE, $U_0$ and $C_{\mathrm{V}}$). The latter point is remarkable when considering the simplicity of KRR, being just a linear expansion of property in training set, and HDAD, being just histograms of distances, angles, dihedrals. Using BR and EN is not recommended if accuracy is of importance, both regressors perform worse across all properties. Apart from predicting highest fundamental vibrational frequency best, RF based models deliver rather unsatisfactory performance. The ECPF4 based models have shown poor general performance in comparison to all other representations studied; it is not recommended for investigations of molecular properties.

We should caution the reader that all our results refer to equilibrium structures of a set of only $\sim$131 k organic molecules. While $\sim$131k molecules might seem sufficiently large to be representative, this number is dwarfed in comparison to chemical space, i.e. the space populated by all theoretically stable molecules, estimated to exceed $10^{60}$ for medium sized organic molecules [156]. Furthermore, ML models for predicting properties of molecules in non-equilibrium or strained configurations might require substantially more training data. This point is also of relevance because some of the highly accurate models described herewithin (MG based) currently use bond based graph connectivity in addition to distance, raising questions about the applicability to reactive processes.

In summary, for the organic molecules studied, we have collected numerical evidence which suggests that the out-of-sample error of ML is consistently better than estimated DFT at B3LYP level accuracy. While this is no guarantee that ML models would reach same error levels if more accurate, explicitly electron correlated or experimental reference data was used, previous studies indicate that similar performance can be expected when using higher levels of theory [31].

More specifically, one might intuitively expect that going beyond hybrid DFT to higher qual-

ity data (either wavefunction based-QM or experiment) in terms of reference methods would represent a more challenging learning problem, and therefore imply the need for larger training set sizes. Results in Ref. [31], however, suggest that ML models can predict the differences between HF and MP2, CCSD, and CCSD(T) equally well using the same training set.

As such, we conclude that future reference datasets for training state-of-the-art ML models of molecular properties should preferably use reference levels of theory which go beyond DFT at B3LYP level of accuracy. While it seems unlikely that for each class of molecules, hundreds of thousands of experimental training data points will become available in the foreseeable future, it might well be possible to reach such scale using efficient implementations of explicit electron correlated methods within high-performance computing campaigns. Finally, we note that future work could deal with improving representations and regressors, with the goal of reaching similar predictive power using less data.

**Table 3.3:** RMSE on out-of-sample data of all representations for all regressors and properties at ∼118k (90%) training set size. Regressors include linear regression with elastic net regularization (EN), Bayesian ridge regression (BR), random forest (RF), kernel ridge regression (KRR) and molecular graphs based neural networks (GG/GC). Additionally, the table contains the mean RMSE of representations for each property and regressor and normalized (by RMSD) mean RMSE (NRMSE) over all properties for each regressor/representation combination.

| | | $U_0$ | $\varepsilon_{\text{HOMO}}$ | $\varepsilon_{\text{LUMO}}$ | $\Delta\varepsilon$ | $\mu$ | $\alpha$ | ZPVE | $C_{\text{v}}$ | $\omega_1$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | eV | eV | eV | eV | Debye | $\text{Bohr}^3$ | eV | cal/molK | $\text{cm}^{-1}$ | |
| EN | CM | 1.28 | 0.459 | 0.78 | 0.903 | 1.19 | 2.28 | 0.0366 | 1.17 | 166.0 | 0.451 |
| | BOB | 0.782 | 0.373 | 0.654 | 0.772 | 1.1 | 2.03 | 0.0292 | 0.904 | 101.0 | 0.372 |
| | BAML | 1.82 | 0.749 | 0.717 | 0.916 | 1.28 | 9.63 | 0.0229 | 1.63 | 91.7 | 0.569 |
| | ECFP4 | 4.99 | 0.295 | 0.45 | 0.497 | 0.971 | 4.76 | 0.354 | 2.08 | 105.0 | 0.482 |
| | HDAD | 0.149 | 0.186 | 0.316 | 0.369 | 0.763 | 1.16 | 0.008 92 | 0.133 | 117.0 | 0.203 |
| | HD | 0.264 | 0.27 | 0.39 | 0.469 | 0.973 | 1.4 | 0.0124 | 0.269 | 126.0 | 0.256 |
| | MARAD | 0.246 | 0.29 | 0.393 | 0.5 | 1.01 | 1.49 | 0.0105 | 0.267 | 130.0 | 0.277 |
| | Mean | 1.36 | 0.375 | 0.529 | 0.632 | 1.04 | 3.25 | 0.0678 | 0.922 | 120.0 | |
| BR | CM | 1.28 | 0.459 | 0.781 | 0.904 | 1.19 | 2.28 | 0.0366 | 1.17 | 167.0 | 0.451 |
| | BOB | 0.764 | 0.368 | 0.652 | 0.771 | 1.09 | 1.97 | 0.028 | 0.884 | 101.0 | 0.365 |
| | BAML | 1.31 | 0.439 | 0.643 | 0.842 | 1.34 | 9.81 | 0.042 | 1.78 | 83.4 | 0.525 |
| | ECFP4 | 4.98 | 0.295 | 0.451 | 0.497 | 0.971 | 4.75 | 0.354 | 2.08 | 105.0 | 0.481 |
| | HDAD | 0.0985 | 0.186 | 0.316 | 0.368 | 0.765 | 1.16 | 0.004 37 | 0.15 | 117.0 | 0.202 |
| | HD | 0.243 | 0.27 | 0.389 | 0.467 | 0.973 | 1.39 | 0.009 14 | 0.256 | 126.0 | 0.255 |
| | MARAD | 0.223 | 0.241 | 0.337 | 0.412 | 0.896 | 1.3 | 0.0109 | 0.257 | 125.0 | 0.245 |
| | Mean | 1.27 | 0.323 | 0.51 | 0.609 | 1.03 | 3.23 | 0.0693 | 0.939 | 118.0 | |
| RF | CM | 0.609 | 0.289 | 0.442 | 0.526 | 0.928 | 1.85 | 0.0264 | 1.04 | 35.2 | 0.28 |
| | BOB | 0.377 | 0.169 | 0.206 | 0.239 | 0.694 | 1.36 | 0.0176 | 0.666 | 6.27 | 0.172 |
| | BAML | 0.399 | 0.156 | 0.179 | 0.209 | 0.668 | 1.41 | 0.0208 | 0.667 | 4.91 | 0.171 |
| | ECFP4 | 5.24 | 0.209 | 0.227 | 0.25 | 0.715 | 5.33 | 0.34 | 2.27 | 26.1 | 0.396 |
| | HDAD | 2.08 | 0.172 | 0.21 | 0.236 | 0.692 | 2.66 | 0.0805 | 1.26 | 6.37 | 0.229 |
| | HD | 2.0 | 0.18 | 0.208 | 0.221 | 0.69 | 2.59 | 0.0761 | 1.22 | 8.08 | 0.225 |
| | MARAD | 0.324 | 0.248 | 0.348 | 0.435 | 0.913 | 1.48 | 0.0147 | 0.446 | 34.4 | 0.234 |
| | Mean | 1.58 | 0.203 | 0.26 | 0.302 | 0.757 | 2.38 | 0.0822 | 1.08 | 17.3 | |
| KRR | CM | 0.185 | 0.181 | 0.245 | 0.309 | 0.664 | 1.14 | 0.006 82 | 0.161 | 49.5 | 0.159 |
| | BOB | 0.0969 | 0.129 | 0.165 | 0.204 | 0.612 | 0.965 | 0.005 01 | 0.122 | 22.9 | 0.117 |
| | BAML | 0.075 | 0.126 | 0.162 | 0.204 | 0.644 | 0.996 | 0.004 41 | 0.111 | 30.6 | 0.122 |
| | ECFP4 | 5.46 | 0.18 | 0.187 | 0.249 | 0.701 | 5.33 | 0.32 | 2.37 | 37.2 | 0.395 |
| | HDAD | 0.0631 | 0.093 | 0.12 | 0.151 | 0.484 | 0.826 | 0.0029 | 0.116 | 36.8 | 0.0985 |
| | HD | 0.0937 | 0.121 | 0.156 | 0.198 | 0.523 | 0.956 | 0.0043 | 0.117 | 33.9 | 0.112 |
| | MARAD | 0.0741 | 0.137 | 0.165 | 0.217 | 0.66 | 1.03 | 0.003 95 | 0.101 | 34.6 | 0.130 |
| | Mean | 0.864 | 0.138 | 0.172 | 0.219 | 0.612 | 1.61 | 0.0496 | 0.442 | 35.1 | |
| GG | MG | 0.307 | 0.0867 | 0.103 | 0.146 | 0.382 | 0.288 | 0.021 | 0.148 | 13.0 | 0.0801 |
| GC | MG | 0.217 | 0.0766 | 0.0926 | 0.119 | 0.145 | 0.342 | 0.017 | 0.133 | 9.87 | 0.0565 |

**Table 4.4:** MAE on out-of-sample data of all representations for all regressors and properties at ∼118k (90%) training set size. Regressors include linear regression with elastic net regularization (EN), Bayesian ridge regression (BR), random forest (RF), kernel ridge regression (KRR) and molecular graphs based neural networks (GG/GC). The best combination for each property are highlighted in bold. Additionally, the table contains mean MAE of representations for each property and regressor; and normalized, by MAD (See Table 4.5), mean MAE (NMMAE) over all properties for each regressor/representation combination.

| | | $U_0$ | $\varepsilon_{\text{HOMO}}$ | $\varepsilon_{\text{LUMO}}$ | $\Delta\varepsilon$ | $\mu$ | $\alpha$ | ZPVE | $C_{\text{v}}$ | $\omega_1$ | NMMAE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | eV | eV | eV | eV | Debye | Bohr$^3$ | eV | cal/molK | cm$^{-1}$ | arb. u. |
| EN | CM | 0.911 | 0.338 | 0.631 | 0.722 | 0.844 | 1.33 | 0.0265 | 0.906 | 131.0 | 0.423 |
| | BOB | 0.602 | 0.283 | 0.521 | 0.614 | 0.763 | 1.2 | 0.0232 | 0.7 | 81.4 | 0.35 |
| | BAML | 0.212 | 0.186 | 0.275 | 0.339 | 0.686 | 0.793 | 0.0129 | 0.439 | 60.4 | 0.231 |
| | ECFP4 | 3.68 | 0.224 | 0.344 | 0.383 | 0.737 | 3.45 | 0.27 | 1.51 | 86.6 | 0.462 |
| | HDAD | 0.0983 | 0.139 | 0.238 | 0.278 | 0.563 | 0.437 | 0.006 47 | 0.0876 | 94.2 | 0.183 |
| | HD | 0.192 | 0.203 | 0.299 | 0.36 | 0.705 | 0.638 | 0.009 49 | 0.195 | 104.0 | 0.236 |
| | MARAD | 0.183 | 0.222 | 0.305 | 0.391 | 0.707 | 0.698 | 0.008 08 | 0.206 | 108.0 | 0.256 |
| | Mean | 0.84 | 0.228 | 0.373 | 0.441 | 0.715 | 1.22 | 0.0509 | 0.578 | 95.1 | |
| BR | CM | 0.911 | 0.338 | 0.632 | 0.723 | 0.844 | 1.33 | 0.0265 | 0.907 | 131.0 | 0.424 |
| | BOB | 0.586 | 0.279 | 0.521 | 0.614 | 0.761 | 1.14 | 0.0222 | 0.684 | 80.9 | 0.343 |
| | BAML | 0.202 | 0.183 | 0.275 | 0.339 | 0.685 | 0.785 | 0.0129 | 0.444 | 60.4 | 0.229 |
| | ECFP4 | 3.69 | 0.224 | 0.344 | 0.383 | 0.737 | 3.45 | 0.27 | 1.51 | 86.7 | 0.462 |
| | HDAD | 0.0614 | 0.14 | 0.238 | 0.278 | 0.565 | 0.43 | 0.003 18 | 0.0787 | 94.8 | 0.182 |
| | HD | 0.171 | 0.203 | 0.298 | 0.359 | 0.705 | 0.633 | 0.006 93 | 0.19 | 104.0 | 0.235 |
| | MARAD | 0.171 | 0.184 | 0.257 | 0.315 | 0.647 | 0.533 | 0.008 54 | 0.201 | 103.0 | 0.226 |
| | Mean | 0.828 | 0.221 | 0.367 | 0.43 | 0.706 | 1.19 | 0.05 | 0.574 | 94.5 | |
| RF | CM | 0.431 | 0.208 | 0.302 | 0.373 | 0.608 | 1.04 | 0.0199 | 0.777 | 13.2 | 0.239 |
| | BOB | 0.202 | 0.12 | 0.137 | 0.164 | 0.45 | 0.623 | 0.0111 | 0.443 | 3.55 | 0.142 |
| | BAML | 0.2 | 0.107 | 0.118 | 0.141 | 0.434 | 0.638 | 0.0132 | 0.451 | 2.71 | 0.141 |
| | ECFP4 | 3.66 | 0.143 | 0.145 | 0.166 | 0.483 | 3.7 | 0.242 | 1.57 | 14.7 | 0.349 |
| | HDAD | 1.44 | 0.116 | 0.136 | 0.156 | 0.454 | 1.71 | 0.0525 | 0.895 | 3.45 | 0.198 |
| | HD | 1.39 | 0.126 | 0.139 | 0.15 | 0.457 | 1.66 | 0.0497 | 0.879 | 4.18 | 0.197 |
| | MARAD | 0.21 | 0.178 | 0.243 | 0.311 | 0.607 | 0.676 | 0.0102 | 0.311 | 19.4 | 0.199 |
| | Mean | 1.08 | 0.142 | 0.174 | 0.209 | 0.499 | 1.43 | 0.0569 | 0.761 | 8.74 | |
| KRR | CM | 0.128 | 0.133 | 0.183 | 0.229 | 0.449 | 0.433 | 0.0048 | 0.118 | 33.5 | 0.136 |
| | BOB | 0.0667 | 0.0948 | 0.122 | 0.148 | 0.423 | 0.298 | 0.003 64 | 0.0917 | 13.2 | 0.0981 |
| | BAML | 0.0519 | 0.0946 | 0.121 | 0.152 | 0.46 | 0.301 | 0.003 31 | 0.082 | 19.9 | 0.105 |
| | ECFP4 | 4.25 | 0.124 | 0.133 | 0.174 | 0.49 | 4.17 | 0.248 | 1.84 | 26.7 | 0.383 |
| | HDAD | 0.0251 | 0.0662 | 0.0842 | 0.107 | 0.334 | 0.175 | 0.001 91 | 0.0441 | 23.1 | 0.0768 |
| | HD | 0.0644 | 0.0874 | 0.113 | 0.143 | 0.364 | 0.299 | 0.003 16 | 0.0844 | 21.3 | 0.0935 |
| | MARAD | 0.0529 | 0.103 | 0.124 | 0.163 | 0.468 | 0.343 | 0.003 01 | 0.0758 | 21.3 | 0.112 |
| | Mean | 0.662 | 0.101 | 0.126 | 0.159 | 0.427 | 0.859 | 0.0383 | 0.333 | 22.7 | |
| GG | MG | 0.0421 | 0.0567 | 0.0628 | 0.0877 | 0.247 | 0.161 | 0.004 31 | 0.0837 | 6.22 | 0.0602 |
| GC | MG | 0.15 | 0.0549 | 0.062 | 0.0869 | 0.101 | 0.232 | 0.009 66 | 0.097 | 4.76 | **0.0494** |

**Table 4.5:** Mean and mean absolute deviation (MAD) for all properties in the QM9 data set, as well as target MAE, and DFT (at B3LYP level of theory) MAE relative to experiment for each property, and the number of molecules used to estimate the values (In parentheses of DFT row). The target accuracies taken from Ref. [83]. Target accuracy for energies of atomization, and orbital energies were set to 1 kcal/mol, which is generally accepted as (or close to) chemical accuracy within the chemistry community. Target accuracies used for $\mu$ and $\alpha$ are 0.1 D and 0.1 Bohr$^3$ respectively, which is within the error of CCSD relative to experiments[133]. Target accuracies used for $\omega_1$ and ZPVE are 10 cm$^{-1}$, which is slightly larger than CCSD(T) error for predicting frequencies [155]. Target accuracies used for $C_\mathrm{v}$ were not explained in article [83]. Section 4.3.3 discusses how the errors for DFT where obtained.

| | $U_0$ | $\varepsilon_\mathrm{HOMO}$ | $\varepsilon_\mathrm{LUMO}$ | $\Delta\varepsilon$ | $\mu$ | $\alpha$ | ZPVE | $C_\mathrm{v}$ | $\omega_1$ |
|---|---|---|---|---|---|---|---|---|---|
| | eV | eV | eV | eV | Debye | Bohr$^3$ | eV | cal/molK | cm$^{-1}$ |
| Mean | −76.6 | −6.54 | 0.322 | 6.86 | 2.67 | 75.3 | 4.06 | 31.6 | 3500 |
| MAD | 8.19 | 0.439 | 1.05 | 1.07 | 1.17 | 6.29 | 0.717 | 3.21 | 238 |
| Target | 0.043 | 0.043 | 0.043 | 0.043 | 0.10 | 0.10 | 0.0012 | 0.050 | 10 |
| DFT | 0.10(69) | NA | NA | NA | 0.10(49) | 0.4(49) | 0.0097(41) | 0.34(16) | 28(41) |

**Table 4.6:** Slopes of the LCs in Fig. 4.3, determined by a linear regression of the two models with largest training set size in each LC for all representations for all regressors and properties. The slopes are estimated under the assumption that the error asymptotically follow power-law decay $\beta N^\alpha$ with training set size ($N$) number of training samples, where $\alpha$ would be the slope.

| | $U_0$ | $\varepsilon_\mathrm{HOMO}$ | $\varepsilon_\mathrm{LUMO}$ | $\Delta\varepsilon$ | $\mu$ | $\alpha$ | ZPVE | $C_\mathrm{v}$ | $\omega_1$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| EN | CM | −0.04 | −0.01 | 0.03 | 0.0 | 0.01 | 0.0 | 0.01 | 0.02 | 0.0 |
| | BOB | −0.08 | −0.01 | 0.0 | −0.01 | 0.03 | −0.01 | −0.02 | −0.03 | −0.03 |
| | BAML | 0.06 | 0.05 | −0.05 | −0.04 | 0.05 | 0.36 | 0.04 | 0.06 | −0.03 |
| | ECFP4 | −0.03 | −0.01 | −0.04 | −0.02 | −0.01 | −0.03 | −0.05 | 0.01 | −0.03 |
| | HDAD | 0.01 | −0.08 | −0.05 | −0.05 | −0.03 | −0.07 | 0.03 | 0.02 | 0.0 |
| | HD | 0.01 | −0.01 | −0.01 | 0.01 | −0.02 | 0.01 | 0.01 | 0.02 | −0.01 |
| | MARAD | −0.01 | 0.01 | −0.04 | −0.02 | −0.03 | 0.03 | 0.0 | −0.01 | −0.02 |
| BR | CM | −0.04 | −0.01 | 0.03 | 0.0 | 0.01 | −0.01 | 0.0 | 0.02 | 0.0 |
| | BOB | −0.07 | 0.0 | 0.0 | −0.02 | 0.03 | 0.0 | −0.03 | −0.04 | −0.05 |
| | BAML | 0.05 | 0.36 | 0.05 | −0.05 | −0.04 | 0.04 | 0.06 | 0.06 | −0.03 |
| | ECFP4 | −0.01 | −0.03 | −0.01 | −0.04 | −0.02 | −0.05 | −0.03 | 0.01 | −0.03 |
| | HDAD | −0.03 | −0.07 | −0.08 | −0.05 | −0.05 | 0.03 | 0.01 | 0.02 | 0.00 |
| | HD | 0.01 | −0.01 | −0.02 | 0.0 | −0.02 | 0.0 | 0.03 | 0.02 | −0.02 |
| | MARAD | 0.38 | −0.02 | −0.02 | −0.03 | −0.02 | −0.02 | 0.14 | 0.3 | −0.04 |
| RF | CM | −0.35 | −0.18 | −0.21 | −0.27 | −0.07 | −0.2 | −0.2 | −0.26 | −0.31 |
| | BOB | −0.48 | −0.24 | −0.3 | −0.26 | −0.08 | −0.27 | −0.43 | −0.45 | −0.19 |
| | BAML | −0.49 | −0.29 | −0.3 | −0.3 | −0.08 | −0.33 | −0.44 | −0.48 | −0.26 |
| | ECFP4 | −0.08 | −0.18 | −0.31 | −0.28 | −0.07 | −0.03 | −0.12 | −0.03 | −0.2 |
| | HDAD | −0.19 | −0.26 | −0.26 | −0.32 | −0.08 | −0.15 | −0.29 | −0.18 | −0.29 |
| | HD | −0.2 | −0.25 | −0.29 | −0.27 | −0.08 | −0.14 | −0.32 | −0.17 | −0.33 |
| | MARAD | −0.57 | −0.19 | −0.24 | −0.29 | −0.03 | −0.18 | −0.3 | −0.4 | −0.25 |
| KRR | CM | −0.36 | −0.25 | −0.33 | −0.33 | −0.22 | −0.36 | −0.36 | −0.39 | −0.26 |
| | BOB | −0.36 | −0.26 | −0.27 | −0.29 | −0.21 | −0.34 | −0.25 | −0.27 | −0.43 |
| | BAML | −0.26 | −0.19 | −0.22 | −0.2 | −0.16 | −0.41 | −0.14 | −0.16 | −0.41 |
| | ECFP4 | 0.9 | −0.17 | −0.28 | −0.2 | −0.06 | 0.72 | 0.51 | 0.95 | −0.22 |
| | HDAD | −0.44 | −0.38 | −0.4 | −0.4 | −0.25 | −0.48 | −0.22 | −0.24 | −0.39 |
| | HD | −0.17 | −0.29 | −0.34 | −0.3 | −0.21 | −0.19 | −0.14 | −0.15 | −0.36 |
| | MARAD | −0.11 | −0.1 | −0.09 | −0.09 | −0.06 | −0.14 | −0.07 | −0.08 | −0.14 |
| GG | MG | −0.36 | −0.32 | −0.35 | −0.34 | −0.26 | −0.31 | −0.23 | −0.35 | −0.36 |
| GC | MG | −0.33 | −0.38 | −0.33 | −0.36 | −0.4 | −0.66 | −0.3 | −0.31 | −0.64 |

**Table 4.7:** Offsets of the LCs in Figs. 4.3, determined by a linear regression of the two models with largest training set size in each LC for all representations for all regressors and properties. The slopes are estimated under the assumption that the error asymptotically follow power-law decay $\beta N^{\alpha}$ with training set size ($N$) number of training samples, where $c$ would be the offset.

| | $U_0$ | $\varepsilon_{\text{HOMO}}$ | $\varepsilon_{\text{LUMO}}$ | $\Delta\varepsilon$ | $\mu$ | $\alpha$ | ZPVE | $C_{\text{v}}$ | $\omega_1$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | eV | eV | eV | eV | Debye | $\text{Bohr}^3$ | eV | cal/molK | $\text{cm}^{-1}$ | |
| EN | CM | 1.41 | 0.379 | 0.421 | 0.701 | 0.757 | 1.39 | 0.0246 | 0.753 | 139 |
| | BOB | 1.59 | 0.333 | 0.548 | 0.73 | 0.553 | 1.38 | 0.0288 | 0.952 | 122 |
| | BAML | 0.0512 | 0.0573 | 0.27 | 0.283 | 0.431 | 0.0126 | 0.0131 | 0.167 | 92.2 |
| | ECFP4 | 5.72 | 0.242 | 0.555 | 0.456 | 0.812 | 5.28 | 0.48 | 1.47 | 121 |
| | HDAD | 0.105 | 0.347 | 0.367 | 0.387 | 0.701 | 0.797 | 0.0208 | 0.135 | 89.6 |
| | HD | 0.165 | 0.217 | 0.352 | 0.336 | 0.915 | 0.581 | 0.008 86 | 0.148 | 121 |
| | MARAD | 0.206 | 0.194 | 0.463 | 0.472 | 0.95 | 0.479 | 0.0083 | 0.239 | 132 |
| BR | CM | 1.4 | 0.38 | 0.423 | 0.705 | 0.746 | 1.5 | 0.0251 | 0.749 | 138 |
| | BOB | 1.25 | 0.291 | 0.545 | 0.749 | 0.531 | 1.19 | 0.0309 | 1.04 | 139 |
| | BAML | 0.102 | 0.105 | 0.5 | 0.554 | 0.4 | 0.0114 | 0.008 35 | 0.222 | 87 |
| | ECFP4 | 5.43 | 0.244 | 0.547 | 0.456 | 0.796 | 5.03 | 0.478 | 1.41 | 120 |
| | HDAD | 0.0568 | 0.366 | 0.409 | 0.477 | 0.779 | 0.94 | 0.002 24 | 0.0637 | 96.8 |
| | HD | 0.145 | 0.221 | 0.368 | 0.369 | 0.917 | 0.634 | 0.004 88 | 0.16 | 124 |
| | MARAD | 0.002 05 | 0.234 | 0.337 | 0.463 | 0.818 | 0.697 | 0.001 73 | 0.005 93 | 157 |
| RF | CM | 26.9 | 1.73 | 3.35 | 8.66 | 1.31 | 10.8 | 0.212 | 16.7 | 470 |
| | BOB | 57.1 | 2.03 | 4.45 | 3.51 | 1.11 | 13.8 | 1.59 | 82.9 | 31.5 |
| | BAML | 59.8 | 3.17 | 4.09 | 4.83 | 1.14 | 28.4 | 2.19 | 121.0 | 58.2 |
| | ECFP4 | 9.31 | 1.16 | 5.28 | 4.16 | 1.04 | 4.96 | 0.994 | 2.3 | 154 |
| | HDAD | 13.6 | 2.51 | 2.68 | 6.55 | 1.12 | 9.93 | 1.61 | 7.15 | 98.0 |
| | HD | 14.4 | 2.34 | 4.13 | 3.73 | 1.22 | 8.3 | 1.99 | 6.64 | 188 |
| | MARAD | 155.0 | 1.56 | 3.79 | 9.28 | 0.871 | 5.53 | 0.328 | 33.1 | 353 |
| KRR | CM | 8.57 | 2.55 | 8.17 | 11.2 | 6.13 | 29.7 | 0.312 | 10.7 | 739 |
| | BOB | 4.48 | 2.09 | 2.83 | 4.42 | 5.04 | 15.3 | 0.0643 | 2.22 | 1950 |
| | BAML | 1.08 | 0.868 | 1.65 | 1.49 | 3.07 | 35.3 | 0.0173 | 0.501 | 2320 |
| | ECFP4 | 0.000 115 | 0.918 | 3.39 | 1.86 | 1.01 | 0.000 943 | 0.000 622 | 0.000 028 3 | 355 |
| | HDAD | 4.22 | 5.39 | 9.14 | 11.4 | 6.05 | 50.1 | 0.0242 | 0.765 | 2080 |
| | HD | 0.483 | 2.49 | 5.96 | 4.87 | 4.21 | 2.63 | 0.016 | 0.466 | 1400 |
| | MARAD | 0.202 | 0.347 | 0.36 | 0.492 | 0.927 | 1.74 | 0.006 48 | 0.194 | 112 |
| GG | MG | 2.95 | 2.36 | 3.67 | 4.85 | 5.0 | 5.88 | 0.0642 | 5.01 | 411 |
| GC | MG | 6.83 | 4.55 | 3.03 | 6.08 | 11.2 | 521.0 | 0.312 | 3.65 | 8140 |

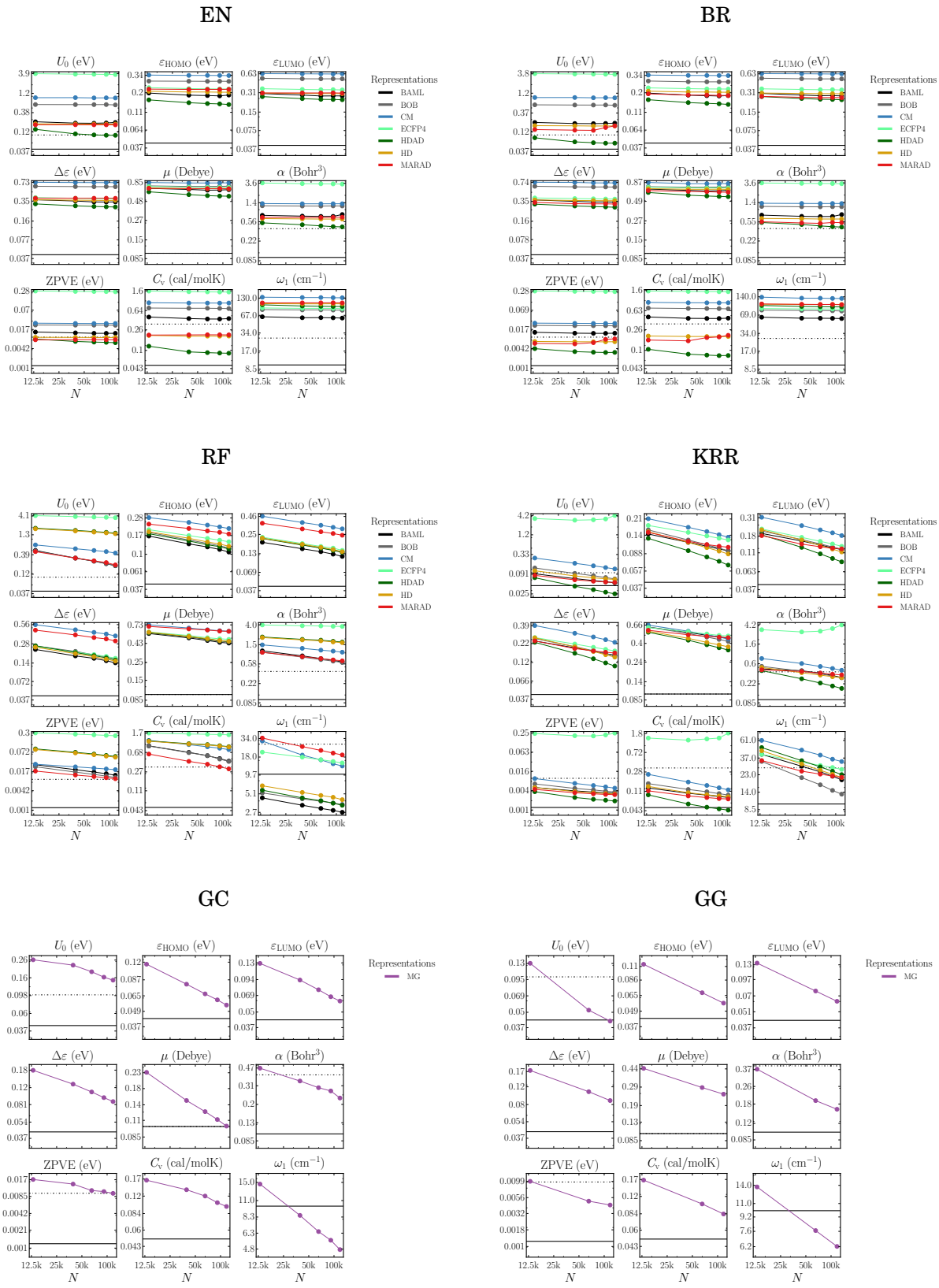



**Figure 4.3:** LCs for combinations of properties, representations, and regressors. Out-of-sample MAE as a function of training set size for QM9 molecules with the property and unit given in the title of each graph. Horizontal solid lines corresponds to target accuracies and horizontal dotted lines corresponds to B3LYP accuracies.

# Chapter 5

# Alchemical and structural distribution based representation for universal quantum machine learning



## 5.1 Executive Summary

We introduce a representation of any atom in any chemical environment for the automatized generation of universal kernel ridge regression-based QML models of electronic properties, trained throughout chemical compound space. The representation is based on Gaussian distribution functions, scaled by power laws and explicitly accounting for structural as well as elemental degrees of freedom. The elemental components help to lower the QML model's LC, and, through interpolation across the periodic table, even enable "alchemical extrapolation" to covalent bonding between elements not part of training. This point is demonstrated for the prediction of covalent binding in single, double, and triple bonds among main-group elements, as well as for atomization energies in organic molecules. We present numerical evidence that resulting QML energy models, after training on a few thousand random training instances, reach chemical accuracy for out-of-sample compounds. Compound data-sets studied include thousands of structurally and compositionally diverse organic molecules, non-covalently bonded protein side-chains, $(H_2O)_{40}$-clusters, and crystalline solids. LCs for QML models also indicate competitive predictive power for various other electronic ground state properties of organic

molecules, calculated with hybrid DFT, including polarizability, heat-capacity, HOMO-LUMO eigenvalues and gap, zero point vibrational energy, dipole moment, and highest vibrational fundamental frequency.

## 5.2 Introduction

Ground-state properties of chemical compounds can generally be estimated with acceptable accuracy using methods such as *ab initio* quantum chemistry or DFT [157]. However, using this information directly to measure similarity results in QML models with rather disappointing predictive power. Alternatively, inductive ML of quantum mechanical properties, i.e. QML models can infer properties directly, or even predict the electron density which in turn can be used to calculate properties [87], by training on a large data sets of reference property/-compound pairs. QML models can have an exceptional trade-off between predictive accuracy and computational cost. For example, in 2017 we showed that QML models can estimate hybrid DFT atomization energies, as well as several other properties, of medium-sized organic molecules with prediction errors lower than chemical accuracy (∼0.04 eV)—multiple orders of magnitude faster than hybrid DFT [18].

The system variable defining the ground-state properties of a given compound is its external potential, a simple function of interatomic distances and nuclear charges. However, when using this information directly to measure similarity results in QML models with rather disappointing predictive power. This can be mitigated by transformation of system variables into "representations". Such transformations can either be designed by human intuition, or be included in the learning problem, e.g. when using NN which include representation learning in the supervised learning task. Letting a NN find the representation has proven to yield models with low out-of-sample prediction errors [39–41]. This approach, however, has the drawback that representation and model are intermingled within the NN, making it less amenable to human understanding, interpretation, adaptation, and further improvement. Furthermore, such machine designed representations do not necessarily lead to better QML performance than representations designed by humans (vide infra).

There are many ways of manually encoding the 3D structure and chemical composition of a compound into a suitable representation. For example, we can represent a compound as a list of interatomic potentials [33, 56, 158]. Another approach consists of creating a fingerprint of the compound, transforming internal coordinates into a fixed set of numbers. For example, this can be done by projecting the coordinates on to a set of basis functions [76], or by creating a "fingerprint" from the topology of the structure [57]. Distributions of internal coordinates represent another systematic approach, shown to yield well performing QML models applicable throughout chemical space [159, 160]. Additional use of bags containing angular and dihedral

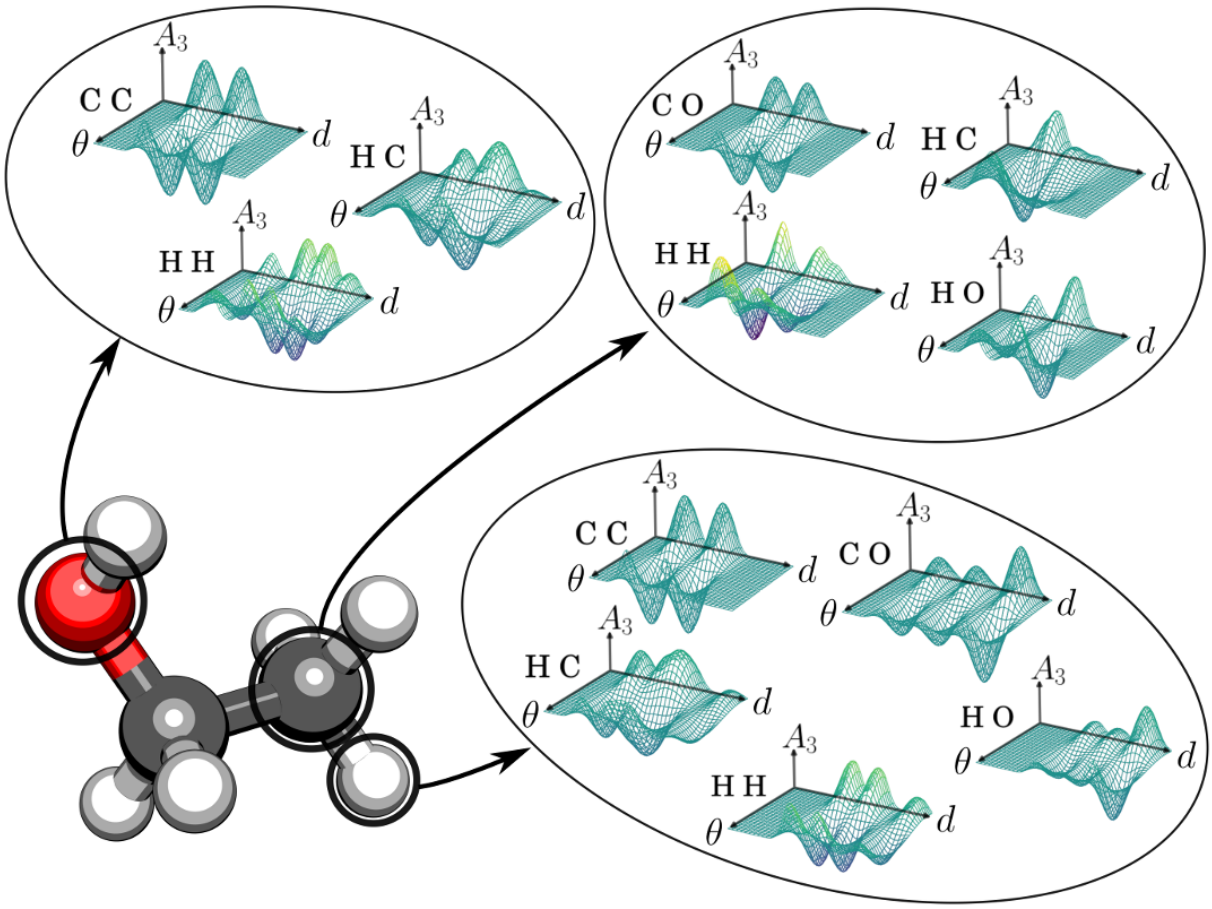



**Figure 2.1:** The three-body term ($A_3$) as a function of radial ($d$) and angular ($\theta$) degrees of freedom in the atomic environments of O, C and H (circled) in ethanol. For simplicity, we show the three-body term without elemental smearing where it reduces to a number of two-dimensional distributions for each element triple.

distributions has led to further improvements in resulting QML models [18, 32, 56, 66]. Bagging based on atom types, however, severely hinders resulting QML models from transferring what has been learned from one atom type to another—a desirable feature for chemically diverse systems.

In this work we introduce a new atomic environment representation, with two key differences to previous distribution-based work. (i) The representation is not binned by atomic types. Instead, compositional information is encoded directly into the distributions. This allows measuring not only structural differences, but also "alchemical" differences between elements in the atomic environments. The idea of computational alchemy, amounting to continuous interpolation of Hamiltonians of two different systems, is well established in quantum chemistry and statistical mechanics and can be exploited for virtual exploration campaigns in chemical space with increased efficiency [161]. Recently, it has been shown that alchemical estimates of covalent bond

potentials can even surpass generalized gradient approximated DFT accuracy [162]. The foundation of a continuous chemical space has been reviewed previously [159]. Alchemical distance measures in the context of QML were already exploited previously when using the Coulomb matrix [33], Fourier series distribution based representations [125], the Faber, Lindmaa, Lilienfeld, Armiento (FLLA) crystal representation [8], and within smooth overlap of atomic potentials (SOAP) representations [26]. For this work we have identified a new functional form with improved performance due to alchemical contributions to the distance measure. (ii) We use a set of multidimensional distributions of interatomic many-body expansions rather than several 1D bins of internal coordinates. The distributions are built recursively, so that an $m$-body distribution contains the same information as the $(m-1)$-body distribution plus additional $m$-body information. This particular combination combines similarity to the potential energy target function and compliance with many known (translational, rotational, permutational) invariances.

## 5.3 Theory

In this section, we first motivate the ideas which have led to this study. Thereafter, we discuss the functional form and the variational degrees of freedom which we have introduced, as well as the resulting compound distances. Then, an analysis of the functional form is performed using the molecule water as an example. Finally, numerical results for parameters optimization runs are discussed.

### 5.3.1 Kernel ridge regression

In order to profit from robustness, ease of error convergence, computational efficiency, and simplicity, we base our studies preferably on KRR models [22–25]. However, we consider this rather a question of taste, and believe that other regressors, such as neural networks, will produce similar results if properly converged.

KRR estimates property $p$ of query compound $\mathbf{C}$ as a weighted sum of kernel basis functions placed on each of $N$ training compounds $\{\mathbf{C}_k\}$,

$$p^{est}(\mathbf{C}) = \sum_{k=1}^{N} \alpha_k K(\mathbf{C}, \mathbf{C}_k), \tag{3.1}$$

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{p}^{\text{train}} \tag{3.2}$$

where the solution for the weights $\{\alpha_k\}$ is obtained through linear regression with regularizer $\lambda$ (typically negligibly small because of absence of noise in training data obtained from quantum chemistry calculations).

Throughout this work we rely on atomistic (scalable) Gaussian kernels,
$K(\mathbf{C}, \mathbf{C}') = \sum_{I \in \mathbf{C}} \sum_{J \in \mathbf{C}'} k(\Delta(\mathcal{A}_M(I), \mathcal{A}_M(J)))$, as already used in [66, 124, 163, 164]. As such, KRR renders the selection of a functional form which represents an atom in its chemical environment mandatory. Obviously, this choice is fundamentally related to our understanding of chemistry, and is known to dramatically affect the performance of resulting QML models, see e.g. [18, 56]. It is for this reason that we draw our inspiration from the fundamental laws of quantum mechanics which specify the definition of system (Hamiltonian) and property (Observable), and which spell out the numerical recipe which links the two [157].

The genesis of this study is due to the fact that the total potential energy, the expectation value of a compound's electronic Hamiltonian, constitutes the central figure of merit for convergence towards the wavefunction by virtue of the variational principle. When considering Eq. (3.2), it should be obvious that the kernel (and thereby representation) is independent of the specific property, units and property dependence is introduced through the regression weights only. This has also already been demonstrated numerically for multiple properties using the same kernel [83]. As such, the role of the kernel is reminiscent of the wavefunction which can be used to predict arbitrarily many observables by evaluating the expectation values of the corresponding operators, always using the same wavefunction: Once the kernel is inverted, arbitrarily many sets of regression coefficients can easily be generated provided that their corresponding property reference values are known.

The Hamiltonian's expectation value, i.e. the potential energy, therefore governs the shape of the wavefunction. We therefore assume that a representation, optimized for energy predictions only, is fundamentally more advantageous than representations obtained by minimizing prediction errors of some integrated observable.

Consequently, the focus employed in this study has been to identify a representation which is inspired by the energy changes occurring due to changes in chemical composition and covalent and non-covalent bonding. The accuracy of quantum mechanics when predicting other properties (observables) as expectation values of operators depends crucially on the quality of the wavefunction. Here, we follow a similar argument: The better the representation the better the energy prediction, implying that energy prediction errors should be minimized in the functional space of the representation.

### 5.3.2 Representation

We use a set of interatomic $M$-body expansions $\mathcal{A}_M(I) = \{A_1(I), A_2(I),$
$A_3(I), \ldots, A_M(I)\}$ which contain up to $M$-body interactions to represent the structural and chemical environment of an atom $I$ in compound $\mathbf{C}$. $A_m(I)$ is a weighted sum that runs over all $m$-body interactions. Each element in the sums consists of Gaussian basis functions, placed on structural and elemental degrees of freedom, and multiplied by a scaling function $\xi_m$. Structural values encode geometrical information about the system, such as interatomic distances or angles. As elemental parameters we use the period $P$ and group $G$ from the periodic table. The scaling functions $\xi_m$ are used to weigh the importance of each Gaussian, based on internal system coordinates. We now consider only the first three distributions in $\mathcal{A}_M(I)$ for an atom $I$. We have also derived, implemented and tested the 4-body $A_4(I)$ distributions. However, the predictive accuracy improvements of resulting QML models were found to be negligible in comparison to the 3-body expansion. Also, the computational cost for generating large kernel matrices increases substantially when going from third to fourth order terms.
The first-order expansion $A_1(I)$ accounts for chemical composition (stoichiometry) and is modeled by a Gaussian function placed on period $P_I$ and group $G_I$ in the periodic table of element $I$:

$$A_1(I) \;=\; \mathcal{N}(\mathbf{x}_I^{(1)}) \;=\; e^{-\frac{(P_I-\chi_1)^2}{2\sigma_P^2}-\frac{(G_I-\chi_2)^2}{2\sigma_G^2}} \tag{3.3}$$

where $\mathbf{x}_I^{(1)} = \{P_I, \sigma_P; G_I, \sigma_G\}$, with respective widths $\sigma_P$ and $\sigma_G$. $\sigma_P$ and $\sigma_G$ can be seen as elemental smearing parameters, which control the near-sightedness of elements in the periodic table. $\chi_1$ and $\chi_2$ represent dummy variables for period and group, to be integrated out when evaluating the Euclidean distance (see Eq. (3.5)). For $A_1(I)$, the scaling function is set to unity, since stoichiometry is geometry independent. We are not aware of other representations in the literature which employ similar distribution functions in the periodic table.
$A_2(I)$ is a product of $A_1(I)$ and a sum that runs over all neighboring atoms $i$:
$A_2(I) = \mathcal{N}(\mathbf{x}_I^{(1)}) \sum_{i\neq I} \mathcal{N}(\mathbf{x}_{iI}^{(2)})\xi_2(d_{iI})$, $\mathbf{x}_{iI}^{(2)} = \{d_{iI}, \sigma_d; P_i, \sigma_P; G_i, \sigma_G\}$, where $d_{iI}$ and $\sigma_d$ correspond to the interatomic distance at which a Gaussian is placed, and its width, respectively. $\xi_2$ corresponds to the 2-body, interatomic distance dependent, scaling function which takes the form of the power laws discussed below. Note that letting $\sigma_P$ and $\sigma_G$ approach zero is equivalent to using a radial distribution function (RDF) for each element pair. This attribute of the representation holds for any of $A_m(I)$, i.e. $\sigma_P, \sigma_G \to 0$ is equivalent to creating a separate distribution for each chemical element $m-$tuple in $A_m(I)$. $A_3(I)$ is the logical extension from $A_2(I)$, it has

a different scaling function with an additional summation, running over all neighboring atoms $j$: $A_3(I) = \mathcal{N}(\mathbf{x}^{(1)}) \sum_{i \neq I} \mathcal{N}(\mathbf{x}_{iI}^{(2)}) \sum_{j \neq i,I} \mathcal{N}(\mathbf{x}_{ijI}^{(3)}) \xi_3(d_{iI}, d_{jI}, \theta_{ij}^I)$, $\mathbf{x}_{ijI}^{(3)} = \{\theta_{ij}^I, \sigma_\theta; P_j, \sigma_P; G_j, \sigma_G\}$. $P_j$ and $G_j$, similarly to $P_i$ and $G_i$, corresponds to the period and group of atom $j$. Again, $\xi_3(d_{iI}, d_{jI}, \theta_{ij}^I)$ is the (three-body) scaling function, and $\theta_{ij}^I$ the principal angle between the two distance vectors $\mathbf{r}_{Ii}$ and $\mathbf{r}_{Ij}$ which span from $I$ to $i$ and $I$ to $j$, respectively. $\sigma_\theta$ is the width of the Gaussian placed at $\theta_{ij}^I$. Letting $\sigma_d$ go to infinity in $A_3$ is equivalent to using a type of angular distribution function (ADF), which in one form or another has already been used in several representations [18, 32, 66]. $A_3$ can therefore be seen as a generalized ADF containing more structural information. Fig. 2.1 illustrates how $A_3(I)$ looks for a hydrogen, carbon, and the oxygen atom in ethanol.

The four body distribution, $A_4(I)$, is defined in eq. 3.4. $A_4(I)$ contains an additional sum over $A_3(I)$, running over all neighboring atoms $k$. $\omega_{ijk}^I$ is the principal angle between the plane spanned by the distance vectors $\mathbf{r}_{Ii}$ and $\mathbf{r}_{Ij}$ and the distance vector $\mathbf{r}_{kI}$. $\sigma_\omega$ corresponds to the Gaussian width at $\omega_{ijk}^I$.

$$\begin{aligned} A_4(I) =& \mathcal{N}(\mathbf{x}^{(1)}) \sum_{i \neq I} \mathcal{N}(\mathbf{x}_{iI}^{(2)}) \sum_{j \neq i,I} \mathcal{N}(\mathbf{x}_{ijI}^{(3)}) \sum_{k \neq j,i,I} \mathcal{N}(\omega_{ijk}^I, \sigma_\omega; P_k, \sigma_P; G_k, \sigma_G) \\ & \xi_4(d_{iI}, d_{jI}, d_{k,I}, \theta_{ij}^I, \omega_{ijk}^I) \end{aligned} \tag{3.4}$$

The scaling functions $\xi$ we have chosen for this work correspond to simple power laws. They have been modified from the leading order two- and three-body dispersion laws by London, $1/r^6$, and Axilrod-Teller-Muto [165, 166], $1/r^9$. Such dispersion expressions were previously already used by some of us [66]. Our scaling functions, however, use different exponents for the radial decay, and set the $C_6$ and $C_9$ coefficients to unity, as early tests indicated better performance for this choice. For periodic systems, however, a very large cutoff radius would be needed in order to converge the distances between two atomic environments, when using the optimized exponents. We have therefore augmented the scaling functions by a previously used soft cutoff function [167], which goes to zero at 9 Å.

### 5.3.3 Distances and scalar products

In order to train and evaluate the KRR model in Eq. (3.1), proper distance measures must be specified. We have found good performance when using a weighted sum of the distances between each $m$-body expansion $\Delta(\mathcal{A}_M(I), \mathcal{A}_M(J))^2 \equiv \sum_{m=0}^{M} \beta_m \Delta(A_m(I), A_m(J))^2$ as a distance between two atomic environments $\mathcal{A}_M(I)$ and $\mathcal{A}_M(J)$. Here, $\beta_m$ is another hyperparameter,

which weighs the importance of each expansion order.

The distances between each distribution term are evaluated as Euclidean ($L_2$) norms, as shown in Eq.3.5. $\varsigma_m$ are normalization constants, which ensures that all individual basis functions integrate to 1 in the $L_2$-norm. All integrals can be solved analytically since they consist of a sum of Gaussian products. The explicit form of the $A_m$ integrals for $m = 1 \dots 3$ is shown in Eq. 3.6.

$$\Delta(A_m(I), A_m(J))^2 = \frac{1}{\varsigma_m^2} \int_{\mathbb{R}^{3m-1}} d\chi_1 \cdots d\chi_{3m-1} (A_m(I) - A_m(J))^2 \tag{3.5}$$

$$\begin{aligned}
\frac{1}{\varsigma_1^2} \int_{\mathbb{R}^2} d\chi_1 d\chi_2 A_1(I) A_1(J) =& \frac{1}{2} \exp(-\frac{(P_I - P_J)^2}{4\sigma_P^2} - \frac{(G_I - G_J)^2}{4\sigma_G^2}) \\
\frac{1}{\varsigma_2^2} \int_{\mathbb{R}^5} d\chi_1 \cdots d\chi_5 A_2(I) A_2(J) =& \frac{1}{2\sqrt{2}} \exp(-\frac{(P_I - P_J)^2}{4\sigma_P^2} - \frac{(G_I - G_J)^2}{4\sigma_G^2}) \\
& \sum_{i \neq I}^{n_I} \xi_2(d_{iI}) \sum_{j \neq J}^{n_J} \exp(-\frac{(d_{jJ} - d_{iI})^2}{4\sigma_d^2} - \frac{(P_i - P_j)^2}{4\sigma_P^2} - \frac{(G_i - G_j)^2}{4\sigma_G^2}) \xi_2(d_{jJ}) \\
\frac{1}{\varsigma_3^2} \int_{\mathbb{R}^8} d\chi_1 \cdots d\chi_8 A_3(I) A_3(J) =& \frac{1}{16} \exp(-\frac{(P_I - P_J)^2}{4\sigma_P^2} - \frac{(G_I - G_J)^2}{4\sigma_G^2}) \\
& \sum_{i \neq I}^{n_I} \sum_{j \neq J}^{n_J} \exp(-\frac{(d_{jJ} - d_{iI})^2}{4\sigma_d^2} - \frac{(P_i - P_j)^2}{4\sigma_P^2} - \frac{(G_i - G_j)^2}{4\sigma_G^2}) \\
& \sum_{k \neq i,I}^{n_I} \xi_2(d_{iI}, d_{kI}, \theta_{ik}^I) \sum_{l \neq j,J}^{n_J} \exp(-\frac{(\theta_{ik}^I - \theta_{jl}^J)^2}{4\sigma_\theta^2} - \frac{(P_k - P_l)^2}{4\sigma_P^2} - \frac{(G_k - G_l)^2}{4\sigma_G^2}) \\
& \xi_3(d_{jJ}, d_{lJ}, \theta_{jk}^J) \\
\frac{1}{\varsigma_2^2} \int_{\mathbb{R}^{11}} A_4(I) A_4(J) d\chi_1 \cdots d\chi_{11} =& \frac{1}{32\sqrt{2}} \exp(-\frac{(P_I - P_J)^2}{4\sigma_P^2} - \frac{(G_I - G_J)^2}{4\sigma_G^2}) \\
& \sum_{i \neq I}^{n_I} \sum_{j \neq J}^{n_J} \exp(-\frac{(d_{jJ} - d_{iI})^2}{4\sigma_d^2} - \frac{(P_i - P_j)^2}{4\sigma_P^2} - \frac{(G_i - G_j)^2}{4\sigma_G^2}) \\
& \sum_{k \neq iI}^{n_I} \sum_{l \neq jJ}^{n_J} \exp(-\frac{(\theta_{ik}^I - \theta_{jl}^J)^2}{4\sigma_\theta^2} - \frac{(P_k - P_l)^2}{4\sigma_P^2} - \frac{(G_k - G_l)^2}{4\sigma_G^2}) \\
\sum_{h \neq k,iI}^{n_I} \xi_4(d_{iI}, d_{jI}, d_{k,J}, \theta_{ij}^I, \omega_{ijh}^I) & \sum_{g \neq l,jJ}^{n_J} \exp(-\frac{(\omega_{ijh}^I - \omega_{jlg}^J)^2}{4\sigma_\omega^2} - \frac{(P_h - P_g)^2}{4\sigma_P^2} - \frac{(G_h - G_g)^2}{4\sigma_G^2}) \\
& \xi_4(d_{jJ}, d_{lJ}, d_{gJ}, \theta_{jl}^I, \omega_{jlg}^I)
\end{aligned} \tag{3.6}$$

However, the third and fourth order terms in the representation are prohibitory expensive to evaluate in practice. The third and fourth order distributions were therefore modified slightly

to solve the angular integrals in Fourier space. The Gaussians $\mathcal{N}(\theta^I_{i,j}, \sigma_\theta)$ and $\mathcal{N}(\omega^I_{i,j,k}, \sigma_\omega)$ in the third and fourth order distribution are first replaced with $\Theta(\theta^I_{i,j}, \sigma_\theta)$ and $\Theta(\omega^I_{i,j,k}, \sigma_\omega)$, respectively, were $\Theta(x, \sigma) = \frac{1}{N}\sum_{k=-N}^{N} \mathcal{N}(x-2\pi k, \sigma) - \mathcal{N}(x-\pi(2k+1), \sigma)$ consists of $4N$ evenly spaced Gaussians with alternating signs. The Fourier transform of $\Theta(x, \sigma)$ w.r.t. $x$ will converge to a Fourier series expansion as $N \to \infty$, whose coefficients converges at a Gaussian rate (See derivation in appendix A). The scalar product of the angular terms can then be evaluated in Fourier space, see eq. 3.7, using precomputed Fourier coefficients, $\mathcal{C}$ and $\mathcal{S}$. defined in eq. 3.8 and 3.9. One set of coefficients $\mathcal{C}$ and $\mathcal{S}$ is computed for every atomic species $T_I$ and $T_J$ in the molecules $I$ and $J$. Further details about the corresponding equations and derivations can also be found in appendix A.

$$
\begin{aligned}
\frac{1}{\varsigma_2^2}\int_{\mathbb{R}^8} A_3(I)A_3(J)d\chi_1\cdots d\chi_8 =& \frac{1}{16}\exp(-\frac{(P_I-P_J)^2}{4\sigma_P^2} - \frac{(G_I-G_J)^2}{4\sigma_G^2}) \\
& \sum_{i\neq I}^{n_I}\sum_{j\neq J}^{n_J} \exp(-\frac{(d_{jJ}-d_{iI})^2}{4\sigma_d^2} - \frac{(P_i-P_j)^2}{4\sigma_P^2} - \frac{(G_i-G_j)^2}{4\sigma_G^2}) \\
& \sum_{a\in T_I}\sum_{b\in T_J}\left[D_{ab}\sum_{n=1}^{N}\{\mathcal{C}_{3ain}(I)\mathcal{C}_{3bjn}(J) + \mathcal{S}_{3ain}(I)\mathcal{S}_{3bjn}(J)\}\right] \\
\frac{1}{\varsigma_3^2}\int_{\mathbb{R}^{11}} A_4(I)A_4(J)d\chi_1\cdots d\chi_8 =& \frac{1}{32\sqrt{2}}\exp(-\frac{(P_I-P_J)^2}{4\sigma_P^2} - \frac{(G_I-G_J)^2}{4\sigma_G^2}) \\
& \sum_{i\neq I}^{n_I}\sum_{j\neq J}^{n_J} \exp(-\frac{(d_{jJ}-d_{iI})^2}{4\sigma_d^2} - \frac{(P_i-P_j)^2}{4\sigma_P^2} - \frac{(G_i-G_j)^2}{4\sigma_G^2}) \\
& \sum_{a\in T_I}\sum_{b\in T_J}\left[D_{ab}\sum_{n=1}^{N}\{\mathcal{C}_{3ain}(I)\mathcal{C}_{3bjn}(J) + \mathcal{S}_{3ain}(I)\mathcal{S}_{3bjn}(J)\}\right] \\
& \sum_{a'\in T_I}\sum_{b'\in T_J}\left[D_{a'b'}\sum_{n=1}^{N}\{\mathcal{C}_{4a'in}(I)\mathcal{C}_{4b'jn}(J) + \mathcal{S}_{4a'in}(I)\mathcal{S}_{4b'jn}(J)\}\right]
\end{aligned}
\tag{3.7}
$$

$$
\begin{aligned}
\mathcal{S}_{3tin}(I) =& \frac{1}{||a||}\sum_{j\neq iI}^{n_I} s_n^{(t)}(\theta^I_{ij})\xi_3(d_{iI}, d_{jI}, \theta^I_{ij}) \\
\mathcal{C}_{3tin}(I) =& \frac{1}{||a||}\sum_{j\neq iI}^{n_I} c_n^{(t)}(\theta^I_{ij})\xi_3(d_{iI}, d_{jI}, \theta^I_{ij}) \\
\mathcal{S}_{4tin}(I) =& \frac{1}{||a||}\sum_{j\neq iI}^{n_I}\sum_{k\neq j,iI}^{n_I} s_n^{(t)}(\omega^I_{ijk})\xi_4(d_{iI}, d_{jI}, \theta^I_{ij}, \omega^I_{ijk}) \\
\mathcal{C}_{4tin}(I) =& \frac{1}{||a||}\sum_{j\neq iI}^{n_I}\sum_{k\neq j,iI}^{n_I} c_n^{(t)}(\omega^I_{ijk})\xi_4(d_{iI}, d_{jI}, \theta^I_{ij}, \omega^I_{ijk})
\end{aligned}
\tag{3.8}
$$

$$
\begin{aligned}
c_n(x) &= \sigma\sqrt{8\pi}\exp[-\frac{(\sigma n)^2}{2}](\cos[xn] - \cos[(\pi + x)n]), n > 0 \\
s_n(x) &= \sigma\sqrt{8\pi}\exp[-\frac{(\sigma n)^2}{2}](\sin[xn] - \sin[(\pi + x)n]), n > 0
\end{aligned}
\tag{3.9}
$$

### 5.3.4 Comparison to other distribution based representations

Probably the largest difference in how $\mathcal{A}$ represents nuclear configurations, when compared to many of the previously published distribution based representations, lies in the 3-body term (since $A_2$ is a radial distribution function if $\sigma_P$ and $\sigma_G$ go to zero). In this subsection, we highlight the differences between $A_3$, or conventional ADF or RDF for representing the structure of the water molecule.

As ADF, we use $A_3$ with the limit $\sigma_d \to \infty$, and we model RDF by $A_2$. Furthermore, no scaling function ($\xi_2 = \xi_3 = 1$) is used and we let $\sigma_P$ and $\sigma_G$ go to zero, since we only examine how representations distinguish structural differences among different geometries of the water molecule. This results in $A_3$ and ADF being $\sum_{i\neq I}\mathcal{N}(d_{iI}, \sigma_d)\sum_{j\neq i,I}\mathcal{N}(\theta^I_{ij}, \sigma_\theta)$ and $\sum_{i\neq I}\sum_{j\neq i,I}\mathcal{N}(\theta^I_{ij}, \sigma_\theta)$ for each element triple, and RDF being $\sum_{i\neq I}\mathcal{N}(d_{iI}, \sigma_d)$ for each element pair.

Fig. 3.2 shows how the distance measure between two water molecules changes as one distorts the geometry of one of them away from its equilibrium structure. Both, RDF as well as ADF result for oxygen as well as for H in large configurational domains with substantially zero distance to the minimum, implying a severe lack of sensitivity. $A_3$, by contrast produces a qualitatively meaningful picture with a single well defined well around the minimum.

We have also studied the performance for modeling the energy of the water molecule. In Fig. 3.3, the training error for atomization energies is shown for a linear kernel KRR model with $A_3$, ADF, RDF, or RDF + ADF as representations. The linear kernel is used as a difficult test in how far representations can model a nonlinear property, such as the energy, in terms of linear basis functions. The errors are significantly lower when using $A_3$ instead of the other representations, including RDF + ADF. Generally, potential energy surfaces of a three-atom system cannot be decomposed into a sum of functions of only one internal coordinate (internuclear distance $\mathbf{d}$ or angle $\theta$). That is, $E(\mathbf{d}, \theta) \neq E(\mathbf{d}) + E(\theta)$. Using a ADF, RDF or a linear combination of the two however would result precisely in such a model, as well as most force fields. This also explains the relatively large errors for these representations, as well as unreliable performance of pair-wise potentials when it comes to distorted molecules. $A_3$ on the other hand does not decouple distances and angles, and can, by construction, model any three-body potential.

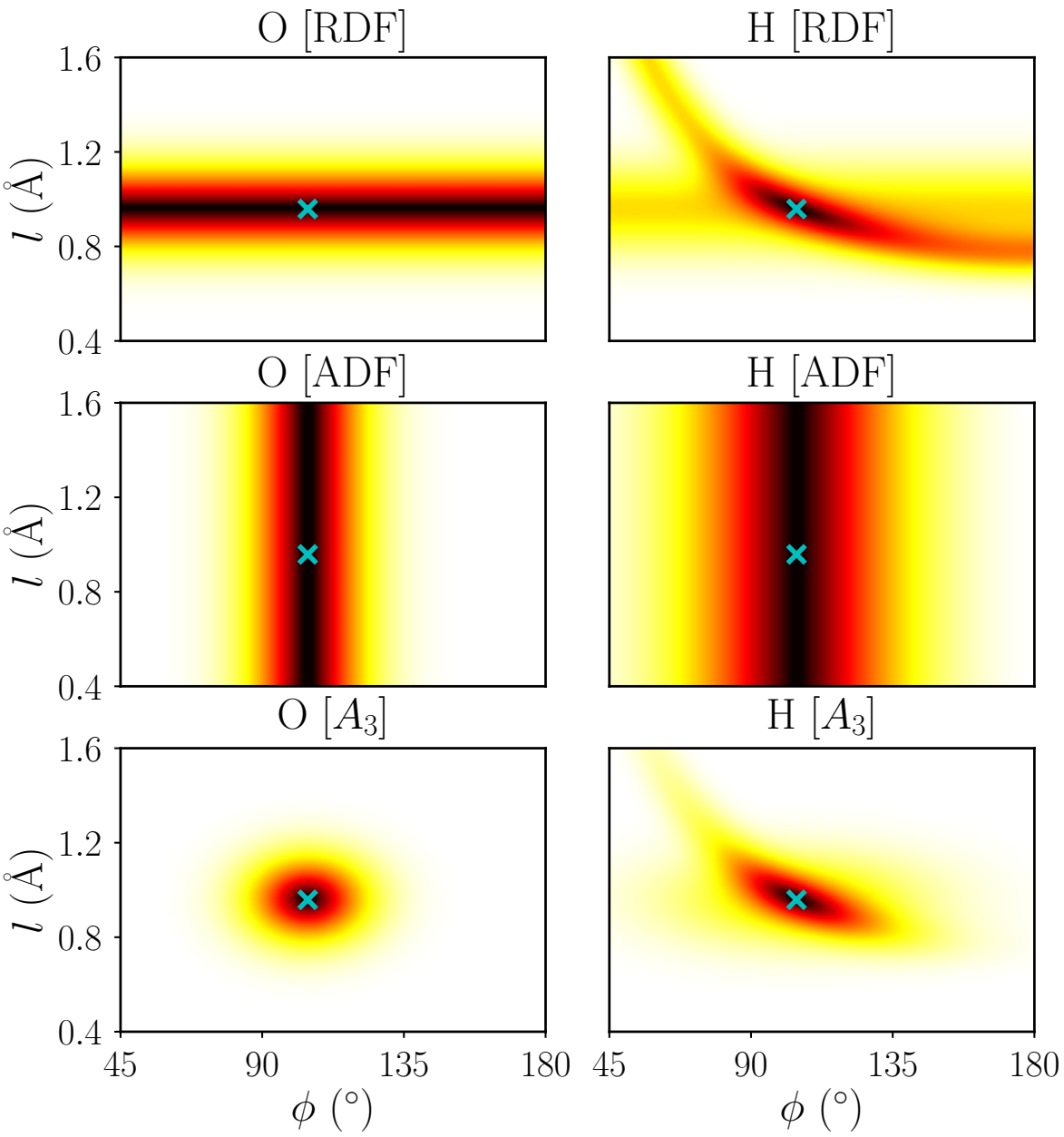



**Figure 3.2:** Heat maps of normalized $L_2$ distances, using three different (not yet scaled) representations (RDF, ADF, and our new representation). The color code from black to white indicates normalized distance, ranging from 0 to 1, respectively. The distances are measured between a reference water molecule, in equilibrium geometry (cross), and a distorted water molecule. Distances are measured separately for the oxygen (LEFT) and hydrogen atoms (RIGHT). The distorted water molecule generated by uniformly stretching of both OH bonds ($d_{\mathrm{OH1}} = d_{\mathrm{OH2}} = l$) and bending the HOH angle ($\phi$) of the reference molecule. The relevant equations for the three representations are given in section 5.3.4.

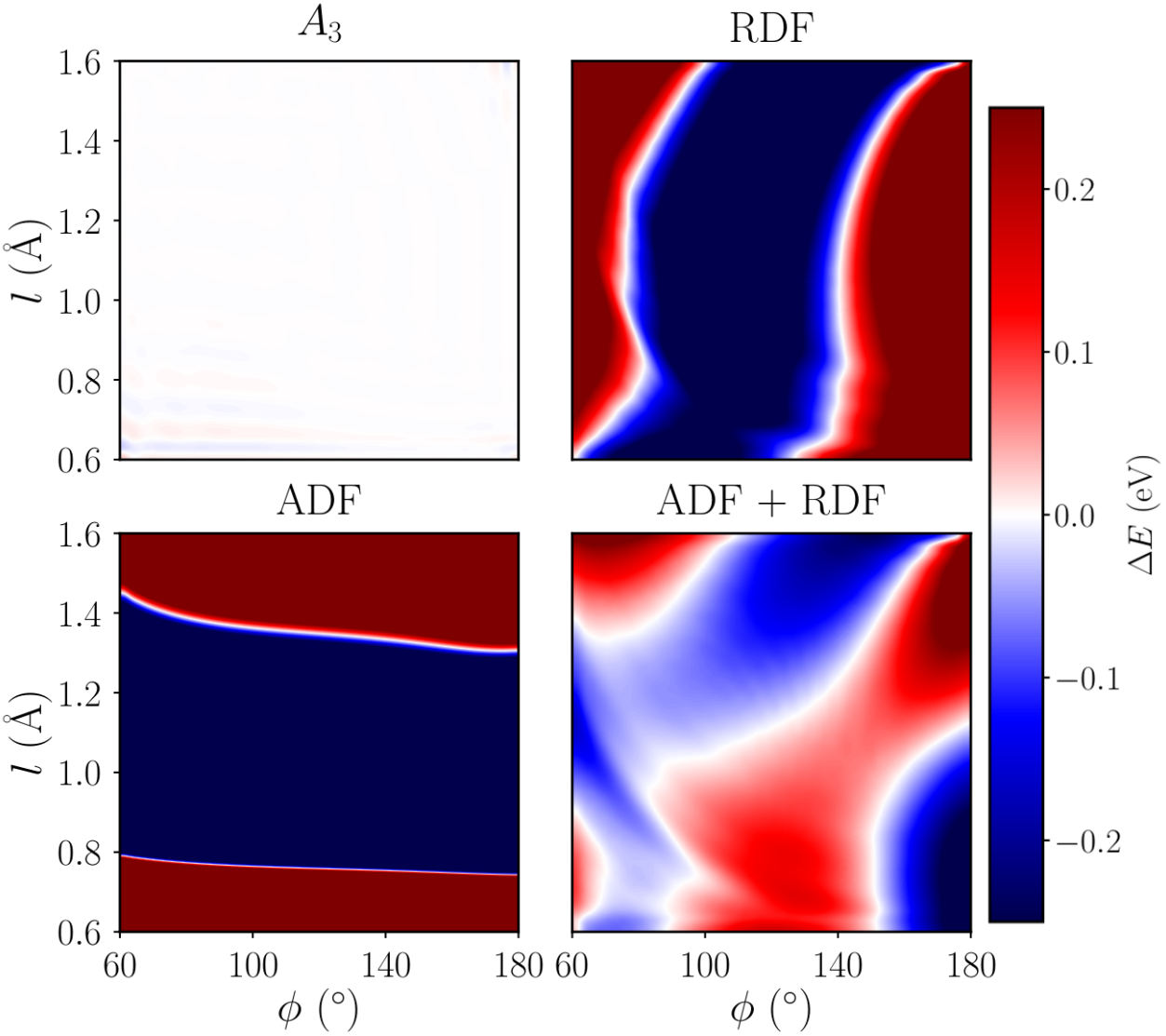



**Figure 3.3:** Heat maps depicting the signed atomization energy error of a water molecule using the same coordinate system as in Fig. 3.2. The errors correspond to linear kernels in KRR fitted to DFT calculated energies (PBE/def2svp) energies. Four representations have been used: TOP LEFT: our new $A_3$ (top left). TOP RIGHT: radial distribution function for each element pair (RDF). BOTTOM LEFT: angular distribution function for each element triple (ADF). BOTTOM RIGHT: RDF + ADF. The training data consists of a equidistant grid of 50-by-50 points along $l$ and $\phi$ within the range of the figures.

These observations give insight as to why our new representation performs better than the other distribution based representations: Using ADF's and RDF's as representations one might be able to capture slices of the many-body picture, the fact that there is a linear mapping between $A_n$ and a $n$-body potential energy surface, however, appears to make it easier to improve the performance also for non-linear kernels.

### 5.3.5 Optimization

#### Hyperparameters

The use of our representation in combination with KRR yields multiple hyperparameters. While one could, in principle, attempt to optimize all of them, using several data sets, and efficient optimizers, such as gradient, Monte Carlo, genetic or simplex methods, we have found that the problem is sensitive only to a small subset of parameters. As such, the exact choice of many hyperparameters is not critical for the out-of-sample errors, and resulting models perform typically well as long as values are used which have similar order of magnitude. Unless otherwise specified, the following hyperparameter values have been used: $\sigma_P = \sigma_G = 1.6$, $\sigma_d = 0.2\,\text{Å}$, $\sigma_\theta = \pi$ , $\beta_1 = 2$, $\beta_2 = \sqrt{8}$, $\beta_3 = 1.6$. For the water cluster and the protein-sidechain-sidechain interaction data set (SSI) there is no to little variation in chemical composition, and no elemental smearing has been used.

#### Scaling power law parameters

We have screened radial exponents $n_2$ and $n_3$ for the scaling functions $\xi_2(d_{iI}) = \dfrac{1}{d_{iI}^{n_2}}$ and $\xi_3(d_{iI}, d_{jI}, \theta_{ij}^I) = \dfrac{1 + 3\cos(\theta_{ij}^I)\cos(\theta_{Ij}^i)\cos(\theta_{iI}^j)}{(d_{iI}d_{jI}d_{ij})^{n_3}}$ , using atomization energies for a subset of the QM9 dataset in order to identify the optimal exponents. Corresponding LCs are shown in Fig. 3.4. First, we have screened $\xi_2$, using $\mathcal{A}_2$ as representation, yielding the lowest off-set for $n_2 = 4$. Keeping this exponent for $\xi_2$ fixed, we then proceeded to screen the exponent $\xi_3$ in $\mathcal{A}_3$ We found that $n_3 = 2$ corresponded to the best exponents for $\xi_3$. We have used these values throughout this work, and unless something else is specified, the optimal scaling functions read,

$$\begin{aligned} \xi_2(d_{iI}) &= \frac{1}{d_{iI}^4} \\ \xi_3(d_{iI}, d_{jI}, \theta_{ij}^I) &= \frac{1 + 3\cos(\theta_{ij}^I)\cos(\theta_{Ij}^i)\cos(\theta_{iI}^j)}{(d_{iI}d_{jI}d_{ij})^2} \end{aligned} \tag{3.10}$$

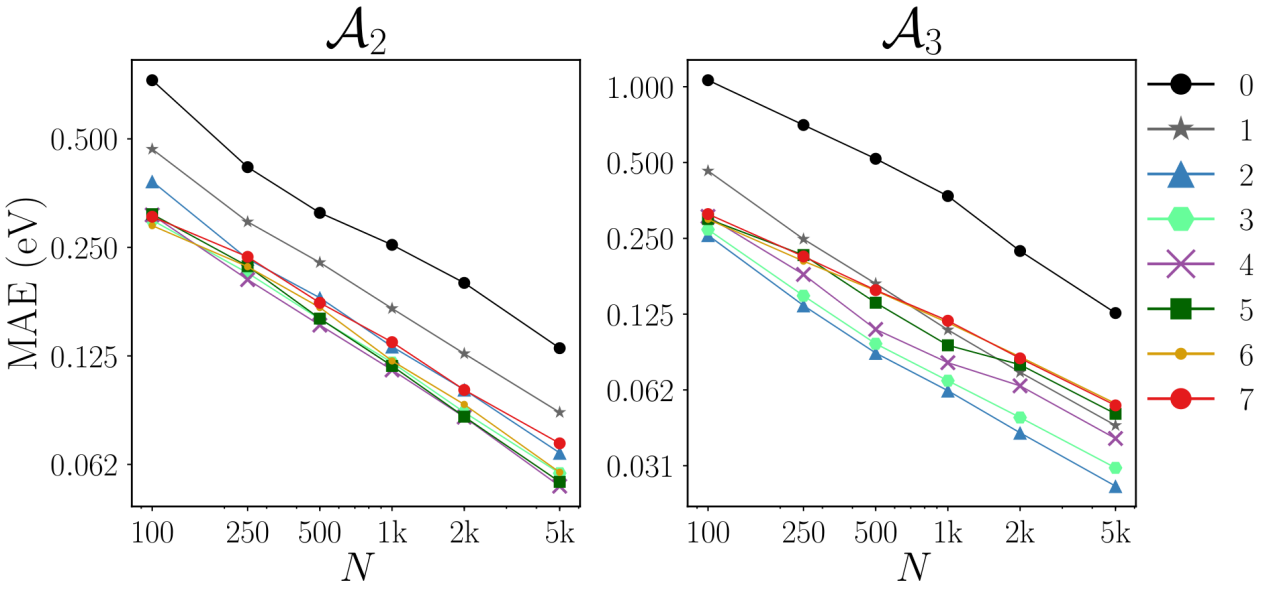



**Figure 3.4:** Optimization of exponents in scaling power laws. LEFT: Out-of-sample MAE for atomization/formation energy predictions as a function of training set size on the QM9 data set. LCs are generated using KRR with $\mathcal{A}_2$ as representation. The legends indicate the exponent $n_2$ used in the scaling power law, $\xi_2(d)$. RIGHT: Out-of-sample MAE for atomization/formation energy predictions as a function of training set size on the QM9 data set. LCs are generated using KRR with $\mathcal{A}_3$ as representation. The legends indicate the exponent $n_3$ used in the scaling power law, $\xi_3(d)$.

#### Alchemical smearing

Parameters associated with the elemental smearing have a strong effect on the predictive power of the QML models. We have screened the corresponding values of $\sigma_P$ and $\sigma_G$ using energy prediction errors for the OQMD and QM9 data set for different training set sizes. These two datasets have been used due to their (relatively) high (OQMD) and low (QM9) chemical diversity in terms of number of differing elements in the the data set. The optimal alchemical Gaussian widths varies only slightly across the two sets, as shown in Fig. 3.5. A circular Gaussian with width $\sigma_P = \sigma_G =\sim$1.6, which amounts to ~90% overlap between neighboring elements, corresponds in a relatively deep well with minimal MAE for the OQMD dataset, no matter the training set size. The fact that the optimal width stays constant with respect to training set size is beneficial: the elemental smearing can be optimized using relatively small training sets, and can then be applied to larger training sets. Comparing the MAE from a model with $\sigma_P = \sigma_G = 0.1$ (which in practice is equivalent zero overlap between different atomic species), using the optimal $\sigma_P = \sigma_G$ lowers the MAE by ~9.9% for the OQMD data set at 100 training samples, which increases up to ~34% when 1k training samples are used. Prediction errors for the QM9 data set indicate similar behavior, yet much less pronounced. For the largest training set (1000 molecules), the optimization well becomes very shallow, consistent

with the lack of compositional diversity in QM9.

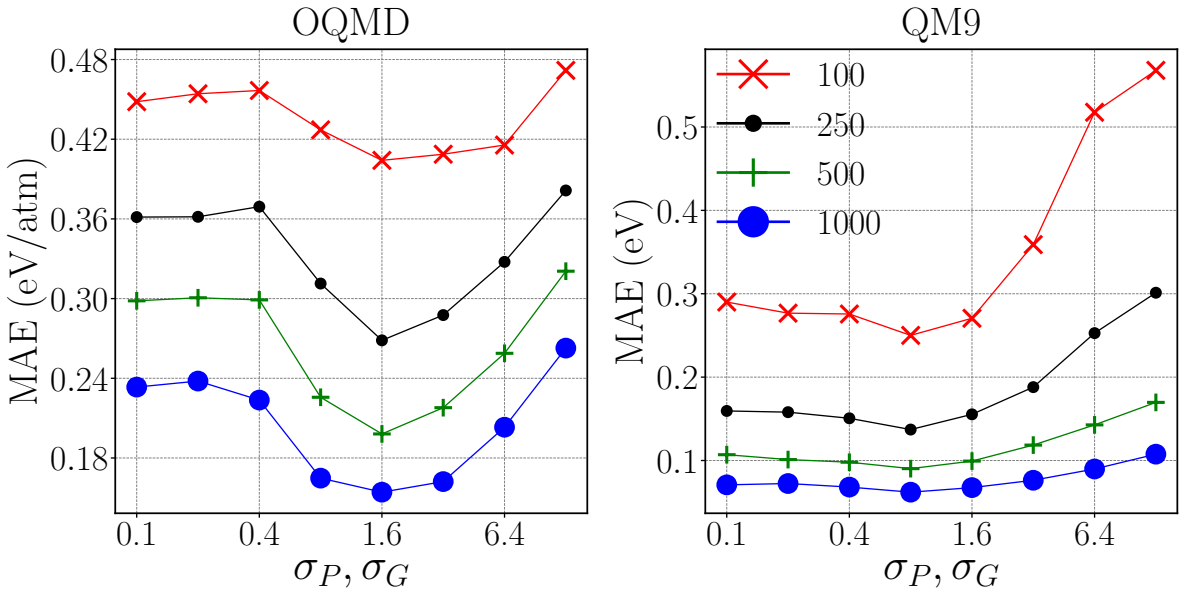



**Figure 3.5:** Changes in out-of-sample MAE as a function of uniform Gaussian width ($\sigma_P$ and $\sigma_G$) used for elemental smearing. Results for energy predictions in the OQMD (LEFT) and QM9 (RIGHT) datasets, respectively. Legends indicate the training set size.

Unsurprisingly, datasets with higher chemical diversity benefit more from using the optimized elemental widths. It may therefore not always be beneficial to include any elemental overlap, especially for datasets with low elemental diversity, as it is computationally more expensive to do so.

## 5.4 Data sets

We have used multiple datasets to benchmark out-of-sample accuracy of energy predictions of our model. These datasets include organic molecules, crystals, biomolecular dimers, water clusters, and main-group diatomics. Some of the datasets are high-quality, have already been published and are in widespread use. Additional low quality data sets have been generated, merely in order to accumulate additional evidence for the relative improvement of the new representation. Further details about the datasets can be found in chapter 3.

Since test set predictions are always close to zero by construction, we exclusively report prediction errors as out-of-sample errors (averaged through cross-validation) with respect to reference validation numbers. All errors reported correspond to at least 10 cross-validation runs for each training set size.

The reader should note that we only report errors of QML models trained on individual datasets.

Simultaneous training on several datasets would introduce significant noise (due to the datasets originating from different levels of theory) and thereby hamper an unequivocal comparative analysis of the results. For future applications and within multi-fidelity QML frameworks [168], pooling the various data sets to train a single QML model might be more desirable.

### 5.4.1 Organic molecules: QM9

The QM9 dataset [63] corresponds to hybrid DFT [90] based structures and properties of 134k organic molecules with up to nine atoms (C, O, N, or F), not counting hydrogen. SMILES strings of these molecules correspond to a subset of the GDB-17 dataset [86]. The 3k organic molecules, which fail SMILES consistency tests [63], were removed before use, i.e. structures where the SMILES strings of the relaxed structure differ from the original GDB-17 SMILES strings.

A random subset of 22k molecules was selected from QM9 for training and testing. 2k molecules were used for testing, and up to 20k for training.

The datasets were sampled differently from QM9 in section 5.5.2 when we investigated how excluding elements from the training set affected the out-of-sample predictions. In total, two tests sets were used, each associated with two training sets. The first test set consisted of molecules containing Nitrogen, and the second test set consisted of molecules containing Oxygen. The two training sets associated with the first test set consisted of molecules containing Nitrogen or not, respectively. The two training sets associated with the second test set consisted of molecules containing Oxygen, or not, respectively.

### 5.4.2 Organic molecules: QM7b

Due to widespread use we also included the more established QM7b dataset [82]. QM7b was also derived from GDB [169]. It contains hybrid DFT (PBE0 [93, 94]) structures and properties of $\sim$7k organic molecules with up to seven atoms (C, O, N, S or Cl), not counting H. We have drawn at random up to 5k molecules for training, and 2k for testing.

### 5.4.3 Biomolecular dimers: SSI

For intra-molecular and non-equilibrium interactions we used a subset of 2356 neutral dimers from recently published SSI dataset [95]. The SSI dataset is a collection of dimers mimicking configurations of interacting amino-acid sidechains as observed in a set of 47 high-resolution

protein crystal structures. The energies correspond to the DW-CCSD(T**)-F12 level of theory [96].

### 5.4.4 Water cluster

We also include a dataset which we calculated for 4'000 structures containing 40 water molecules, drawn from an NVE-molecular dynamics (MD) trajectory of a water droplet, simulated at 300K, treated with the TIP3P potential [170] as implemented in CHARMM C41a1 [171]. For each structure, a single-point energy was calculated at the DFT level using the PBEh-3c method[172]. Additional details, as well as the full dataset, can be found in the SI.

### 5.4.5 Solids: OQMD

We have used the Inorganic Crystal Structure Database [106, 107] subset corresponding to the OQMD by Wolverton and co-workers [104, 105]. This data-set has already been used to develop and benchmark random forest based QML model (Voronoi) [108]. The dataset consists of ∼30k crystal structures and formation energies, calculated using high-throughput DFT with generalized gradient approximation (GGA+U). We have used a random subset consisting of 3k structures with less than 40 atoms in the unit cell and formation energies lower than 5 eV/atom for training and testing. 1k crystals were used for testing, and up to 2k for training.

### 5.4.6 Solids: Elpasolites

We have also tested our representation for the Elpasolite crystal structure data set [8]. This data set consists of ∼ 10k Elpasolite structures and DFT (PBE [116]) formation energies. The crystals correspond to quaternary main group elemental composition with all elements up to Bismuth (39 in total). We have used a random subset of 7k structures, with up to 6k and 1k for training and testing, respectively.

### 5.4.7 Maingroup diatomics

To test the predictive power for alchemical interpolation we have also included a set of previously published DFT (PBE [116]) results for single, double, and triple bonds among main-group diatomics saturated with hydrogens [162]. The training/test splits are generated differently, as explained in Sec. 5.5.2.

## 5.5 Results and Discussion

Using LCs generally resulting in straight lines when recorded on log-log plots due to their inverse power law relationship [52]), we first present numerical results which indicate the predictive power of our QML model for atomization and formation energies in various data sets. When available for the same data set, we also compare to other QML models in the literature. Thereafter, the alchemical extrapolation capacity is demonstrated for predicting covalent bonds in molecules with elements that were not part of training. Finally, log-log plots of LCs for nine electronic ground-state properties of organic molecules (QM9) are reported and discussed.

### 5.5.1 Energies of molecules, clusters, and solids

Fig. 5.6 displays the performance overview for energy predictions on six different data sets (QM9, QM7b, SSI, water, elapsolites, OQMD). Mean absolute out-of-sample energy prediction errors are shown as a function of training set size. The results indicate remarkable performance for all data sets, indicating a well-working QML model yielding systematic improvement with increasing training set size. The LCs also indicate out-of-sample MAEs which are consistently lower, or similar, than previously published models in the literature. For QM9, the MAE reaches the highly coveted chemical accuracy threshold (1 kcal/mol or $\sim 0.043$ eV for enthalpy of formation) with only 2k training points on the QM9 dataset. Previously published QML models had to include an order of magnitude more training molecules to reach such accuracy. This is similar to the amount of training molecules necessary when using the Coulomb matrix representation in conjunction with semi-empirical or DFT based baselines in order to estimate electron correlated energies, as demonstrated in 2015 with the $\Delta$-ML model [31].

For QM9, atomic Spectral London Axilrod-Teller-Muto (aSLATM) [66] and SOAP multi kernel model [26, 173] reach a performance nearly as good as our QML model. aSLATM, however, performs worse for the SSI and the Water cluster. The SOAP multi kernel QML model, however, performs an expansion in kernel function space acting on the distance for which all degrees of freedom have already been integrated out. As such it is, strictly speaking, not the same as as an improved representation, but rather an improved regressor. Note that single kernel based SOAP QML models perform significantly worse. The reader should take notice however that in the SOAP LC results presented in Fig 5.6, the $\sim$3k structures which had failed the SMILES consistency test, were included. As such, theses QML models are not exactly comparable, and the SOAP results are still likely to slightly improve if these faulty structures were to be removed.

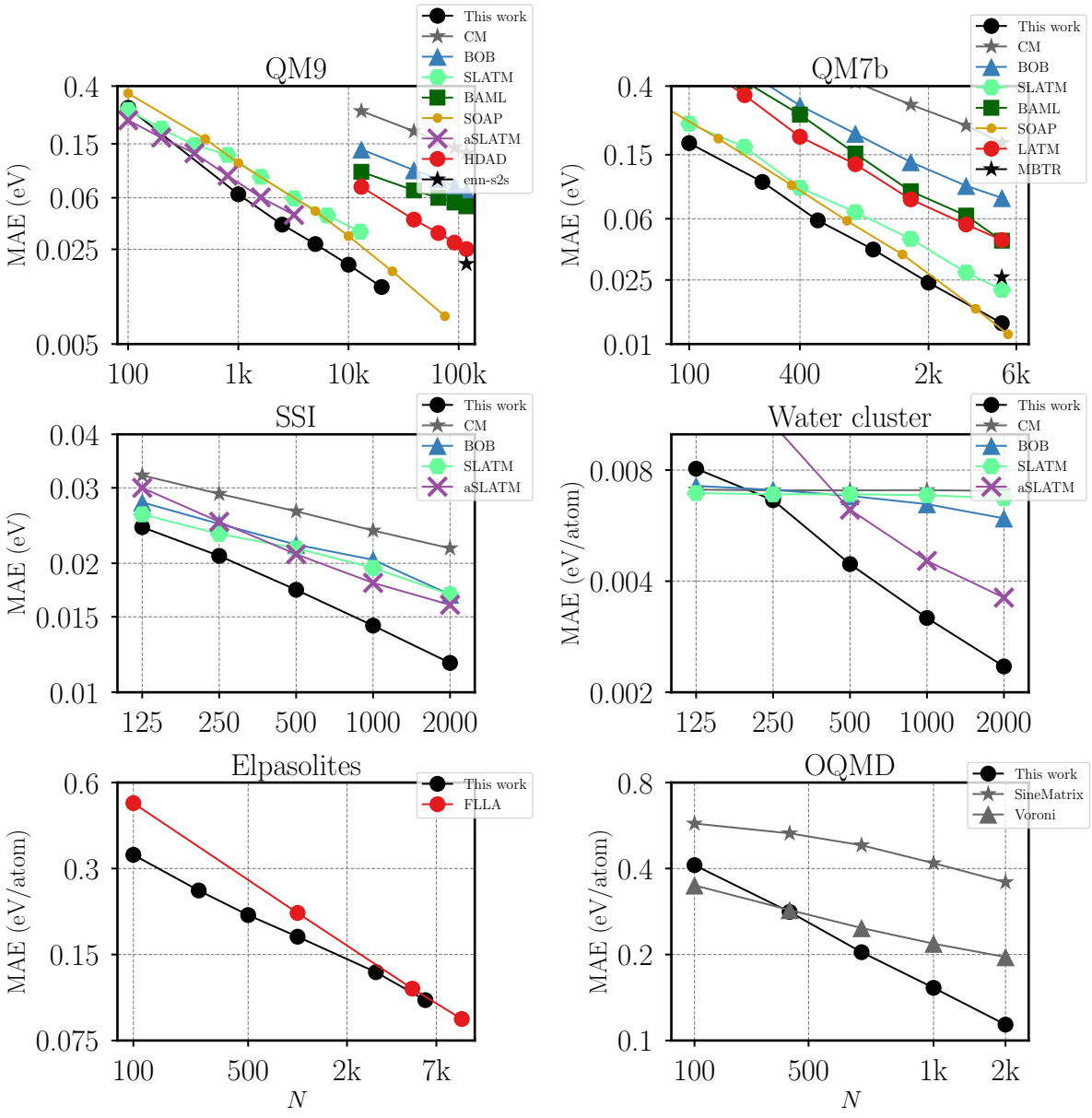



**Figure 5.6:** LCs for atomization/formation energy predictions corresponding to various QML models. Out-of-sample MAE is shown as a function of training set size for molecules (QM9 and QM7b), protein side-chain dimers (SSI), liquid water ($(H_2O)_{40}$ snapshots (Water cluster) and crystal (OQMD and Elpasolites) data-sets.

One should also note that the SOAP results shown for QM7b correspond to the multi-kernel SOAP kernel [29, 173].

Other models presented correspond to Coulomb matrix (CM) [33], bags of bonds (BOB) [158], Bonds and Angles based ML (BAML) [56], Histogram of Distances, Angles, and Dihedrals (HDAD) [18], Spectral London Axilrod-Teller-Muto (SLATM), aSLATM [66], the crystal representation by Faber, Lindmaa, Lilienfeld, Armiento (FLLA) [8], the sine matrix [174], and the many-body tensor representation (MBTR) [32]. We also compared to QML models which are not based on KRR, such as the message passing neural network model (enn-s2s) [39], and a Voronoi-tessellation based random forest model (Voronoi) [108].

The MAE of our new QML model is consistently the lowest for all data sets and large training sets. For the set of 4,000 non-equilibrium water clusters, there is a noticeable difference between the global (CM, BOB and SLATM) and the atomic representations (i.e., aSLATM and the new model we introduce in this work): The global models exhibit very little learning at first, only for larger $N$ the LCs begin to turn downward. The atomic models, however, our new representation based QML model as well as aSLATM, improve rapidly with increasing training data set size. We believe that sorting and crowding in the global representations makes it difficult to accurately account for the purely geometrical changes in structures that contribute to total energy variations.

Impressive predictive power is also observed for the OQMD dataset, a structurally and compositionally very diverse set of solids. Our new model has a lower out-of-sample MAE for all $N$ when compared to the sine matrix representation on the OQMD dataset. The offset of the LC of our new model is larger compared to that of the Voronoi-based random-forest model [108]. However, the learning rate of our QML model is significantly steeper, surpassing the Voronoi model already at just $\sim$ 250 training samples. Results for a solid state variant of the CM, designed for use in periodic systems, has also been included (sine matrix) [174]. It has a similar slope as the Voronoi model, but a substantially larger off-set.

For the elpasolite data set, [8], with large composition diversity but identical crystal structures, the learning-curve of the FLLA representation has a slightly higher off-set than our new QML model, yet exhibits a steeper LC. Our model converges towards the same slope for larger training set sizes. We can only speculate on the reasons for such behavior. The FLLA representation differs qualitatively from the other representations in this study: It does not include any explicit information about coordinates and only encodes periodic row and column of the elements which populate each crystal structure site. The QML model then learns to infer ground state energies

without knowing the exact configuration. This leads to a very low dimensional model that is still unique for the system, which might be the cause of the lower slope. This however needs to be investigated more carefully before any conclusions can be drawn.

### 5.5.2 Alchemical predictions

Our new scaled many-body expansion explicitly accounts not only for distributions of interatomic distances and angles but also for elemental distributions in the periodic table. We have therefore studied its capability to predict covalent binding of molecules containing chemical elements which were not present in the molecules used for training. More specifically, we have investigated single, double, and triple bonds with one bonding atom coming from group (IV), i.e. C, Si, or Ge. In order to increase covalent bond order, we have varied the valency of their bonding partner as follows: For single bonds, group IV atoms are bound to halogens (group VII). For double bonds, group IV atoms atoms are bound to chalcogen atoms (group VI), and for triple bonds, group IV atoms are bound to group V atoms. Dangling valencies of group IV atoms have been saturated with hydrogen. Similar covalent bonding potentials have also recently been used in order to assess the predictive power of first and second order perturbation theory based alchemical predictions [162].

In order to test the alchemical "extrapolation", we trained on the covalent bonds of all other compounds (16 curves) which did contain neither the group IV atom nor the corresponding bonding partner in question. The predictive power for the out-of-sample molecule, on display in Fig. 5.7, is impressive. Albeit not quantitative (chemical accuracy is not reached), the results are semi-quantitative and certainly provide a physically very adequate picture of the covalent bonding in single, double, and triple bonds for main-group atoms in periods 2 to 4. The fact that predictions for the central elements $H_2SiS$ are more accurate (easier to interpolate) than others is consistent with this interpretation. We also note that the deviation is the worst for 2nd-row elements (due to lack of *d*-orbitals they differ substantially more from 3rd and 4th row than 3rd and 4th row differ from each other). Because of poor performance, we do not compare to other representations in this test.

We have also investigated thousands of organic molecules in order to obtain improved quantitative statistics on the question of how out-of-sample prediction errors of different representations are affected when elements in the test set are excluded from training. We tested our new model, with and without elemental smearing ($\sigma_P, \sigma_G = 1.6$ and $\sigma_P, \sigma_G \to 0$ respectively), and we also included BOB and CM representations for comparison. Details about how the training and

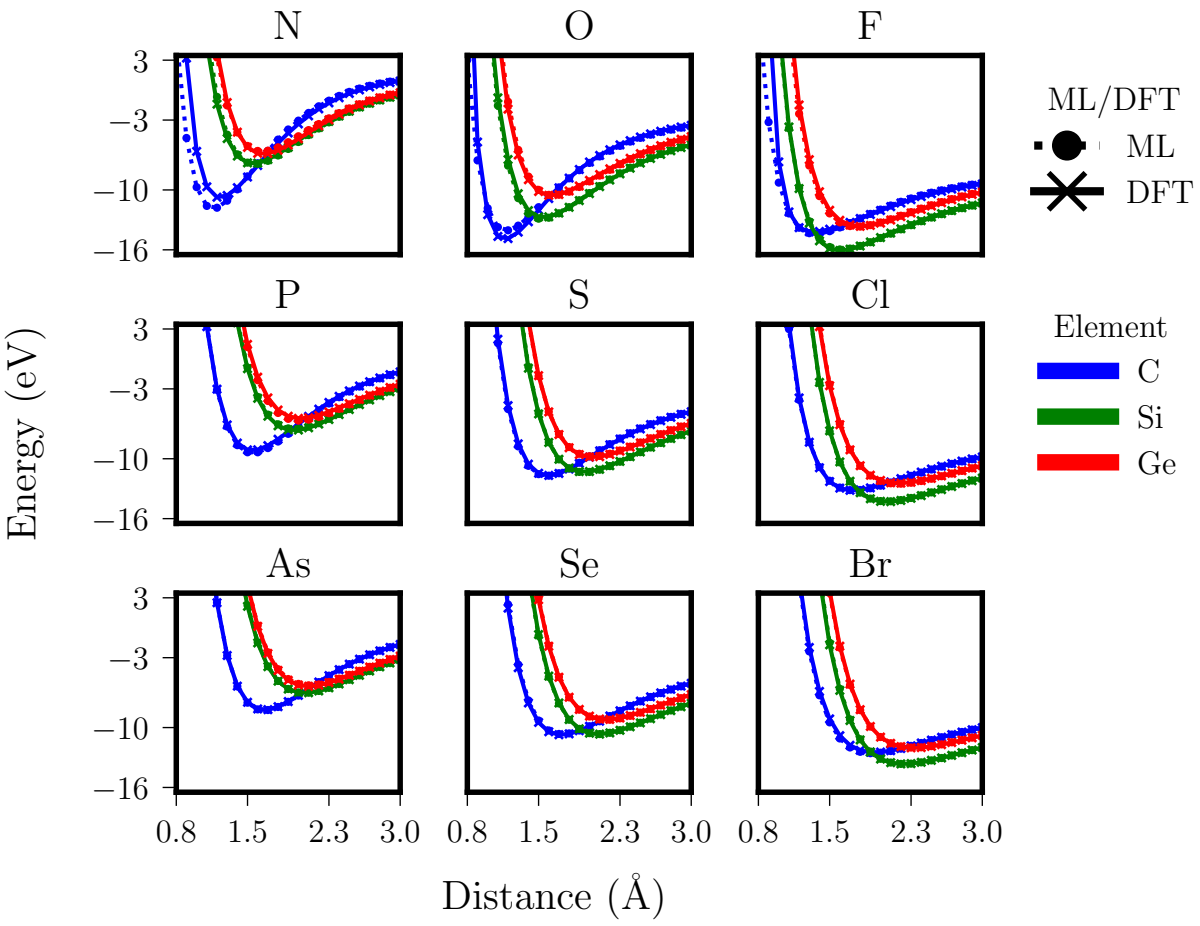



**Figure 5.7:** Covalent bond potentials calculated by DFT (star) and estimated by QML (circle) for 27 main-group diatomic molecules. Bonding occurs between a group IV element (C blue, Si green, or Ge red), and halogens (single bond), chalcogen (double bond), or a group V element (triple bond). Columns correspond to triple (LEFT), double (MID), and single bonds (RIGHT). Rows correspond to the period of the group IV atom's binding partner: 2nd period (TOP), 3rd period (MID), 4th period (BOTTOM).

test sets were selected can be found in section 5.4.1.

Corresponding prediction errors (MAE, RMSE, and Maximal error) for test (training) sets with (without) Nitrogen or with (without) Oxygen are shown in Fig. 5.8. Obviously, the models perform best when both, training *and* test molecules, contain the same elements. However, even for models trained on sets without Nitrogen/Oxygen and without elemental smearing, considerable predictive power and, maybe more importantly, systematic improvement with training set size is observed for our new model. Use of elemental smearing results in a slight improvement.

The loss in accuracy due to absence of elements in training is substantially worse for BOB and CM. Notably, BOB, in general considered to be more accurate than CM [18, 158], experiences a more dramatic loss than CM resulting in BOB being worse than CM. This can

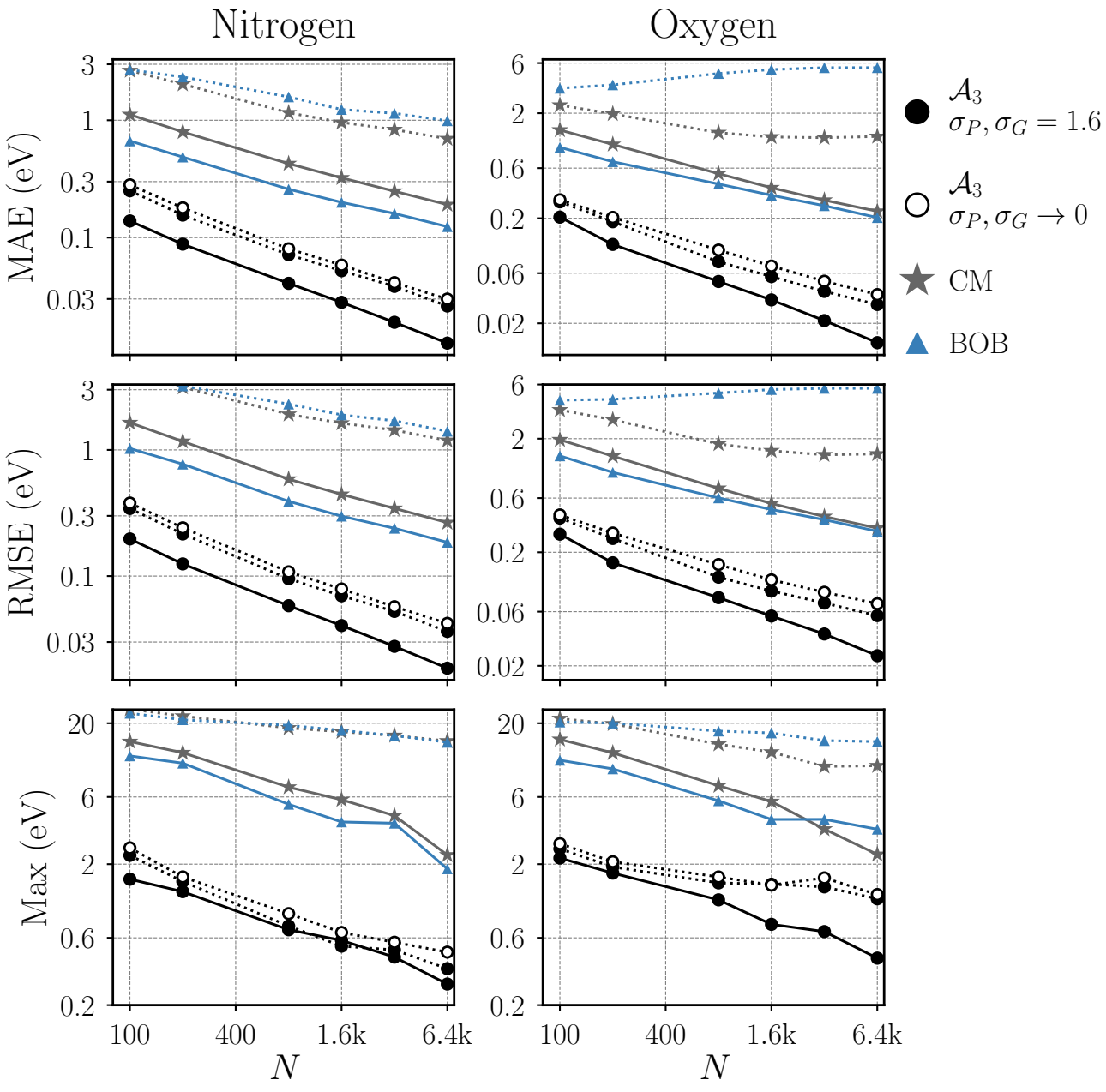



**Figure 5.8:** LCs for atomization energy predictions using training sets with (FULL) and without (DOTTED) elements Nitrogen (LEFT) or Oxygen (RIGHT). All molecules were extracted from QM9 (see section 5.4.1). QML predictions have been made using our new model, with (FILLED circle) and without (EMPTY circle) elemental smearing, as well as KRR with CM and BOB representation (following Ref. [30, 158]).

be understood if one considers the fact that BOB bags nuclear Coulomb repulsion terms by element pairs, whereas the CM matrix directly compares the coulomb interactions without (explicit) regard for which elements-pairs are being compared, effectively already performing an "alchemical" comparison. This allows the CM-based models to meaningfully interpolate even towards elements not part of training. Our model, however, clearly outperforms CM and BOB: For example, its MAE reaches chemical accuracy ($\sim$0.03 eV) for the Nitrogen lacking training set at $\sim$6.4k molecules (rather than at $\sim$1.6k training molecules containing N). Note that CM and BOB based KRR models are far from reaching such accuracy even after being trained on molecules *containing* the element in question!

The relatively high accuracy of our model achieved without any of the elemental smearing, however, is surprising. We would have expected that the lack of the appropriate elements in training introduces prediction errors which can no longer be decreased through the addition of more molecules. However, the LCs in Fig. 5.8 do not indicate any worsening of the learning rate. While possible that the expected deterioration could still be observed for larger training sets, we do find it surprising that such high accuracy can be reached without any elemental smearing at all.

These results clearly demonstrate that alchemical extrapolation is possible when interpolating elemental groups and periods in the periodic table through an appropriate representation. Since the representation is continuous in the corresponding compositional space, we also believe that indication is given that the calculation of alchemical derivatives is meaningful, similar in spirit to Ref. [175].

### 5.5.3 Other ground state properties of molecules

Finally, we also investigated how well QML models, based on our new representation and optimized for energies, perform when predicting other ground-state quantum properties, part of the QM9 dataset. More specifically, we have included atomization energies, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) eigenvalues as well as the HOMO-LUMO gap, dipole moment, polarizability, zero point vibrational energy, heat capacity, and the vibrational frequency of the highest lying fundamental ($\omega$). Results are shown in Fig. 5.9, and provide overwhelming evidence that resulting models enable predictions systematically improving with training set size, no matter what property. For comparison, we have also included results for the aSLATM model. aSLATM results are typically worse when dealing with extensive properties, such as energies, polarizability, or heat-capacity. When deal-

ing with intensive properties, such as eigenvalues or dipolemoment, aSLATM is on par or even slightly better than our model, with the exception of $\omega$. $\omega$ corresponds to the vibrational stretching frequency of CH, NH, or OH bonds, a property with hardly any variance at all. Previously we have seen that this property is best predicted by a random forest model which have relatively poor performance for all other properties [18]. The interpretation is that predicting this property is much more a classification problem, then a supervised learning task.

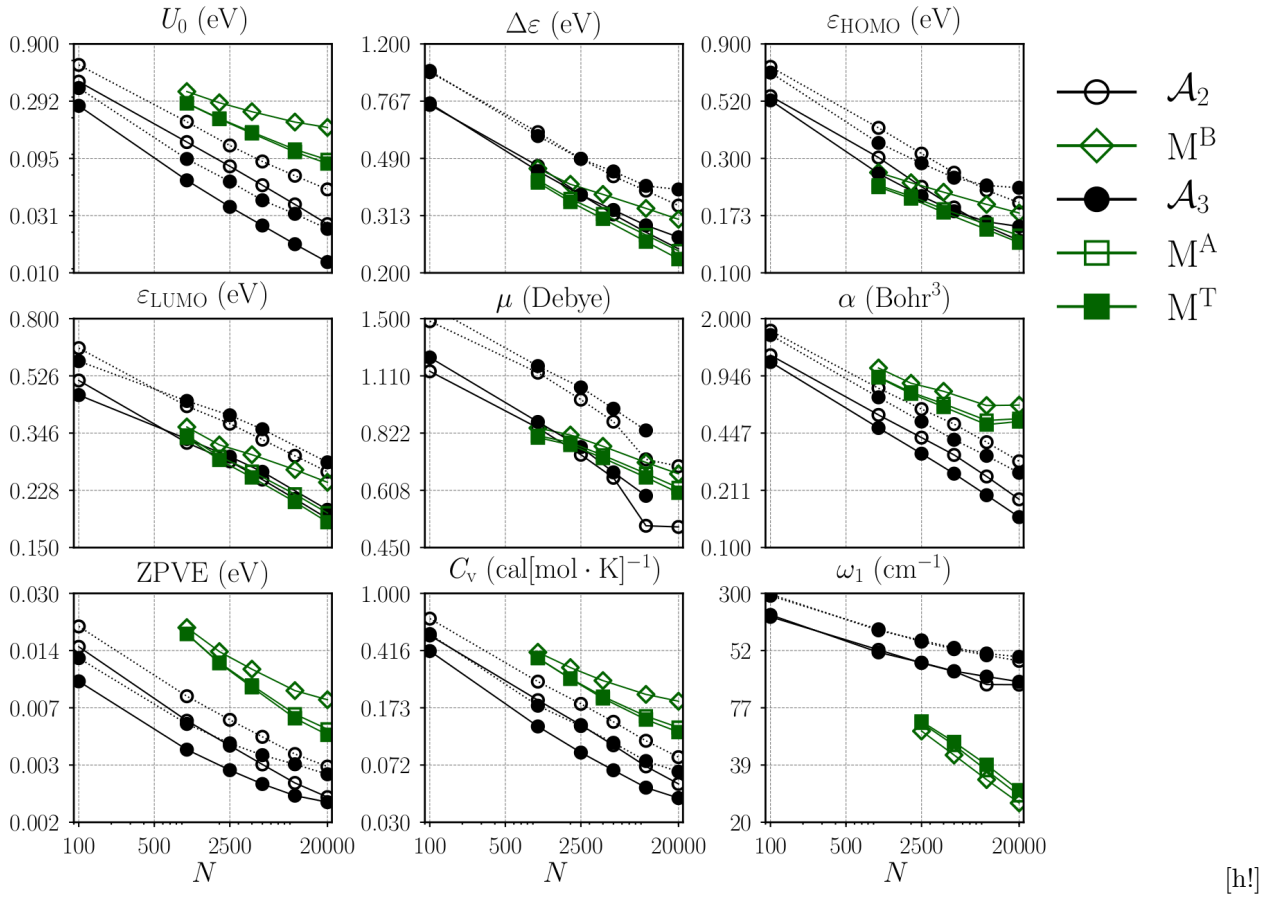
Fig. 5.9 also includes LCs for the RMSE, indicating a slightly higher offset (to be expected) and systematic improvement with training set size with similar slopes as the MAE. This is an assuring result, indicating that also predictions for outliers improve as training set size is increased, as already discussed in Ref. [31, 56].

Furthermore, for Fig. 5.9 we have also distinguished between two and three body contributions (as well as four-body for BAML). For all properties but for $\omega$ the trend meets the expectation, as also already confirmed previously for BAML [56]: Addition of the higher order term systematically lowers the LCs by a significant amount.

## 5.6 Conclusion

We have introduced a universal representation of an atom in a chemical compound for use in QML models. An atom is represented by a sum of multidimensional Gaussians, each term corresponding to elemental, atom-pairwise, and angular distributions and scaled by respective power laws. For the compounds and properties studied we have found four-body contributions to be insignificant.

Most system-independent hyperparameters, such as exponents in scaling functions and Gaussian widths where found to not be critical to the preference of resulting QML-models, as long as "reasonable" heuristics [83] was used. This could, however, be explored more systematically within future work. Analytical expressions have been derived for corresponding distances between arbitrary chemical compounds. These distances can directly be used within KRR based QML models of electronic ground state properties. For energies of organic molecules, water clusters, amino-acid side chains, and crystalline solids the resulting QML models lead to LCs with very low off-set and steep learning rate. For compositionally diverse systems chemical accuracy ($\sim$1 kcal/mol) can now be reached using only thousands of training instances. We have also studied the effect of explicitly accounting for inter-elemental distances in the periodic table: Our new QML model can produce semi-qualitatively accurate covalent bonding potentials for




[h!]

**Figure 5.9:** LCs for out-of-sample MAE (filled lines) and RMSE (dashed lines) as a function of training set size $N$ for nine electronic ground state properties in the QM9 dataset. QML predictions have been made using either a molecular kernel and BAML as representation, or atomic kernels with our new representation. The BAML representation includes bonds ($M^B$); bonds and angles ($M^A$); and bonds, angles and torsional angles ($M^T$). Predicted properties include: atomization energy, at 0 Kelvin ($U_0$); HOMO-LUMO gap ($\Delta\varepsilon$); HOMO eigenvalue ($\varepsilon_{\mathrm{HOMO}}$); LUMO eigenvalue ($\varepsilon_{\mathrm{LUMO}}$); norm of dipole moment ($\mu$); static isotropic polarizability ($\alpha$); zero point vibrational energy (ZPVE); heat capacity at room temperature ($C_{\mathrm{v}}$); and the highest fundamental vibrational frequency ($\omega_1$).

single, double, and triple bonds which include chemical element-pairs which were not part of training. For thousands of organic molecules we also demonstrated that our model, after being trained on molecules which do not contain Nitrogen or Oxygen, still outperforms by a margin CM or BOB based models trained on molecules which do contain Nitrogen or Oxygen. For various electronic ground state properties of organic molecules, numerical results indicate that has remarkable predictive power can be reached.

While the reference data used in this study has mostly been obtained at the hybrid DFT level of theory, the steep LCs of our QML models suggest that it has now become a realistic possibility to obtain a sufficiently large training set at post-Hartree-Fock level of theory (or from experiment), and to use it for the training of QML models which enable subsequent high-throughput screening efforts with similar accuracy.

Combining our new representation with the recently proposed QML model trained on molecular quasi-particles representing atoms-in-molecules ( a.k.a. "am-on" approach) might provide the possibility to generate accurate models which scale with size of query system [66]. Subsequent work will deal with forces and other properties.

# Chapter 6

# Operators in quantum machine learning: Response properties in chemical space



## 6.1 Executive Summary

The role of response operators is well established in quantum mechanics. We investigate their use for universal quantum machine learning models of response properties in molecules. After introducing a theoretical basis, we present and discuss numerical evidence based on measuring the potential energy's response with respect to atomic displacement and to electric fields. Prediction errors for corresponding properties, atomic forces, and dipole moments improve in a systematic fashion with training set size and reach high accuracy for small training sets. Prediction of normal modes and IR-spectra of some small molecules demonstrates the usefulness of this approach for chemistry.

This work was done in collaboration with Anders S. Christensen, who both implemented the models and performed most of the calculations. I derived the models, equations, and helped to analyze the results.

## 6.2 Introduction

Time-independent electronic ground-state quantum properties can be expressed as expectation values of the electronic wave function and an operator, typically defined via the quantum-classical correspondence principle. The performance of supervised machine learning models of these quantum properties, a.k.a. quantum machine learning (QML), [8, 33, 82, 176] can be conveniently assessed using LCs which monitor the decay of the out-of-sample prediction error (deviation of predicted properties from reference for query compounds not included in training) as a function of compound training set size $N$. Due to the leading prediction error decaying as $a/N^b$, log-log plots have become the recommended practice in the field with $\log(a)$ and $b$ denoting the off-set and learning rate (or efficiency), respectively.[17, 52, 53] While in principle, supervised ML models can be generated for any cause and effect relationship, it is the very philosophy of QML that representation (and kernel function when using kernel ridge regression) is property independent[83, 177] in the same way in which the electronic wave function and its Hamiltonian are property independent. However, there is a select and highly relevant set of quantum properties which can be understood as response properties, obtained through the use of response operators and perturbation theory. Common examples include derivatives of the energy with respect to the nuclear displacement or charge, an external electric field, an external magnetic field, or nuclear magnetic moments, and can efficiently be accounted for within density functional theory.[178, 179] We note in passing that energy response properties also form the basis for conceptual density functional theory,[180, 181] as well as computational alchemy.[159, 162, 175, 182–186] It has previously been observed that prediction errors of many conventional quantum machine learning models of response properties can converge relatively slowly, even for machine models that are able to achieve remarkably high accuracy for energies.[18, 19, 68, 82, 83] In this paper we investigate if the use of response operators is beneficial for deriving improved QML models which afford LCs with lower off-sets and better learning rates.

Perhaps the most relevant quantum response property is the force exerted on each atom in the system, the first order energy derivative with respect to nuclear displacement.[187] Quite recently, tremendous efforts have been made to predict atomic forces accurately within QML models for the purpose of running *ab initio* quality molecular dynamics simulations at low computational cost.[28, 29, 38, 61, 73–81] Treating the force as the first derivative of the energy is tantamount to using the gradient operator, as commonly implemented in quantum chemistry

packages. Doing so leads directly to energy conservation, a crucial property for most statistical mechanics applications, which has already also been obtained by others [60, 61]. The use of response operators, however, has not yet been applied generally to generate QML models for other response properties.

Here, we extend the principle of using response operators to investigate the potential total energy and its response to a change in (i) atomic coordinates and (ii) an external electric field, i.e. the dipole moments. Other QML models capable of predicting dipole moments have already been published.[41, 56, 62, 70, 82–85]

The work by Schütt *et al.* presents a neural network that is able to predict the dipole moment of the QM9 dataset[63, 86] with very high accuracy[41] by training on the dipole moment vector itself. Other approaches rely on a charge model predicted from a neural network to estimate intensities in an infrared spectrum where the frequencies are obtained from a molecular dynamics simulation.[62, 84] Similarly to Schütt *et al.*, we propose to learn the dipole moment by training on the quantum mechanical observable directly, but in contrast we train a model to describe the energy for which the dipole moment can be calculated as a response property by taking the derivative of the energy with respect to an external electric field. The modeling of highly accurate molecular potential energy surfaces has also been thoroughly investigated with several ML techniques, due to their important connection to infrared (IR) spectroscopy.[188–191] We show how our operator formalism can lead to ML potential energy surfaces that reproduce the vibrational normal modes of molecules across chemical space and even reproduces the IR spectrum of a molecule by using the relevant response operators with a suitable training set.

This paper is organized as follows: first we present the derivation for a kernel-based regression model capable of predicting response properties by letting the response operator act on the kernels. We then implement a representation that allows us to simultaneously train on properties that depend on both the external electric field as well as the internal degrees of freedom of the molecule. The hydrogen fluoride molecule is used as a toy model to demonstrate the principle. We benchmark the operator-based machine learning model on a number of existing data sets that benchmark forces, energies, and dipole moments across chemical space, and show how our response model improves learning the dipole moment of molecules when compared to conventional kernel ridge regression models. Lastly, we discuss how the model naturally couples force and energy predictions with dipole moment predictions, and we show how the response model can directly predict properties related to second order derivatives, including mixed derivatives,

such as infrared intensities, harmonic vibrational frequencies, and normal modes.

## 6.3 Theory

### 6.3.1 Operator Quantum Machine Learning (OQML)

Within kernel-based regression,[22–25] the total potential energy $U_C^*$ of a query molecule $C$ in its electronic ground-state, can be decomposed into a sum of local energies of its $I$ atomic contributions, which are calculated using a basis of kernel functions:

$$U_C^* = \sum_{I \in C} U_{\text{local}}^* \left(q_I^*\right) = \sum_{I \in C} \sum_J \mathcal{K}\left(q_J, q_I^*\right) \alpha_J \tag{3.1}$$

where $J$ is an atomic environment in the basis, $\alpha_J$ is its regression weight, and $q_I$ is the representation of the $I$'th atom in the molecule, and here the asterisk denotes query atom.
Writing Eq. 3.1 in matrix form, we have:

$$\mathbf{U} = \mathbf{K}\boldsymbol{\alpha} \tag{3.2}$$

Note that in contrast to conventional kernel ridge regression (KRR) and Gaussian Process Regression (GPR) based QML models,[177] this kernel matrix is not symmetric since first dimension is over the atoms used to build the basis and the second dimension has one entry for each observable, e.g. energies for molecules in the example in Eq. 3.2.
In this work, we approximate a response property $\omega$, i.e. an observable which can be computed by applying a differential operator $\mathcal{O}$ acting on the energy $U^*$, defined in Eq. 3.1,

$$\omega = \mathcal{O}[\mathbf{U}] = \mathcal{O}[\mathbf{K}]\boldsymbol{\alpha} \tag{3.3}$$

The set of regression coefficients, $\boldsymbol{\alpha}$, can obtained by minimizing the Lagrangian

$$J(\boldsymbol{\alpha}) = \sum_\gamma \beta_\gamma \|\mathcal{O}_\gamma[\mathbf{U}^{\text{ref}}] - \mathcal{O}_\gamma[\mathbf{K}\boldsymbol{\alpha}]\|^2_{L_2(\Omega_\gamma)} \quad \equiv \tag{3.4}$$

$$\sum_\gamma \beta_\gamma \int_{\Omega_\gamma} \Big[\mathcal{O}_\gamma[\mathbf{U}^{\text{ref}}] - \mathcal{O}_\gamma[\mathbf{K}\boldsymbol{\alpha}]\Big]^T \Big[\mathcal{O}_\gamma[\mathbf{U}^{\text{ref}}] - \mathcal{O}_\gamma[\mathbf{K}\boldsymbol{\alpha}]\Big] \tag{3.5}$$

with respect to $\boldsymbol{\alpha}$ over some training set of known values of $\mathcal{O}[\mathbf{U}^{\text{ref}}]$. $\Omega_\gamma$ is the domain over which the corresponding operator should be minimized, e.g. all rotational degrees of freedom if the operator acts on a SO(3) group. $\gamma$ denotes the specific perturbation (of any order), so

that the model can be trained for multiple properties simultaneously, for example energies, gradients, and dipole moments, and $\beta_\gamma$ is a weight-factor that can be used to adjust weighting in the regression step. For simplicity we pick $\Omega$ such that $\int_\Omega = 1$ for the remainder of this study. $\boldsymbol{\alpha}$ can be obtained e.g. by solving the associated normal equations or using an orthogonal factorization such as a QR[192] or a singular-value decomposition (SVD). The corresponding normal equation (see appendix B for derivation) to this problem is given by

$$\boldsymbol{\alpha} = \Big[\sum_\gamma \beta_\gamma \int_{\Omega_\gamma} \mathcal{O}_\gamma[\mathbf{K}]^T \mathcal{O}_\gamma[\mathbf{K}]\Big]^{-1} \Big[\sum_\gamma \beta_\gamma \int_{\Omega_\gamma} \mathcal{O}_\gamma[\mathbf{U}^{\text{ref}}]^T \mathcal{O}_\gamma[\mathbf{K}]\Big] \tag{3.6}$$

However, solving the normal equations can be numerically unstable since it effectively squares the condition number, i.e. $\kappa(\mathbf{K}^T\mathbf{K}) = (\kappa(\mathbf{K}))^2$.

For the practical implementation and the results discussed in the following, an SVD factorization has been used to solve Eq. 3.4, as it is has several practical and efficient implementations. In contrast to the QR factorization, the SVD factorization is also numerically stable, even if $\mathbf{K}$ is rank-deficient, e.g. if $\mathbf{K}$ contains rows or columns that correspond to atoms or molecules that are identical or only differ by symmetry operations to which the representation is invariant.

In the case of under-determined equations, the SVD factorization is performed ignoring singular values smaller than a threshold, which can be treated as a hyperparameter similarly to regularization within ordinary KRR.

### 6.3.2 Operators

This section is dedicated to discussing some important response operators in quantum mechanics, defining the domain $\Omega$ over which the Lagrangian is to be minimized, and providing the corresponding solutions to the integrals in Eq. 3.6.

We define the response operator for some external parameter $\eta = \{\eta_x, \eta_y, \eta_z\}$ which can be written as $\mathcal{O}_{\delta\eta} \equiv \dfrac{\partial}{\partial\eta}$. Applying such an operator would map the scalar field to a three dimensional vector field. All rotational degrees of freedom can then be integrated out with the following solutions. The solutions to the two integrals in Eq. 3.6, respectively, are thus

$$\int_{\Omega_{\delta\eta}} \mathcal{O}_{\delta\eta}[\mathbf{K}]^T \mathcal{O}_{\delta\eta}[\mathbf{K}] = \frac{1}{3} \sum_{\nu\in x,y,z} \Big(\frac{\partial}{\partial\eta_k}\mathbf{K}\Big)^T \Big(\frac{\partial}{\partial\eta_\nu}\mathbf{K}\Big) \tag{3.7}$$

$$\int_{\Omega_{\delta\eta}} \mathcal{O}_{\delta\eta}[\mathbf{U}^{\text{ref}}]^T \mathcal{O}_{\delta\eta}[\mathbf{K}] = \frac{1}{3} \sum_{\nu\in x,y,z} \Big(\frac{\partial}{\partial\eta_\nu}\mathbf{U}^{\text{ref}}\Big)^T \Big(\frac{\partial}{\partial\eta_\nu}\mathbf{K}\Big). \tag{3.8}$$

Similarly this procedure can be used to solve the equations for the second order response operator, with respect to two different perturbations $\eta$ and $\eta'$:

$$\int_{\Omega_{\delta\eta\delta\eta'}} \mathcal{O}_{\delta\eta\delta\eta'}[\mathbf{K}]^T \mathcal{O}_{\delta\eta\delta\eta'}[\mathbf{K}] = \frac{1}{9} \sum_{\nu,\nu' \in x,y,z} \left(\frac{\partial^2}{\partial\eta_\nu \partial\eta'_{\nu'}} \mathbf{K}\right)^T \left(\frac{\partial^2}{\partial\eta_\nu \partial\eta'_{\nu'}} \mathbf{K}\right) \tag{3.9}$$

$$\int_{\Omega_{\delta\eta\delta\eta'}} \mathcal{O}_{\delta\eta\delta\eta'}[\mathbf{U}^{\mathrm{ref}}]^T \mathcal{O}_{\delta\eta\delta\eta'}[\mathbf{K}] = \frac{1}{9} \sum_{\nu,\nu' \in x,y,z} \left(\frac{\partial^2}{\partial\eta_\nu \partial\eta'_{\nu'}} \mathbf{U}^{\mathrm{ref}}\right)^T \left(\frac{\partial^2}{\partial\eta_\nu \partial\eta'_{\nu'}} \mathbf{K}\right) \tag{3.10}$$

A step-by-step derivation of these equations is given in appendix B. We note that the above equations are only true if the kernel is invariant with respect to rotations around $\theta$ and $\phi$, which is true for the FCHL representation used in conjunction with a rotationally invariant kernel function, such as the Gaussian kernel.

Now we can explicitly write the matrix elements for the operators investigated within this study. In the following, the indices uppercase $I$, $J$, and $K$ correspond to atomic centers, and lowercase $i$, $j$, and $k$ correspond to molecules.

The unperturbed kernel corresponds to the energy or identity operator acting on the kernel. The elements of the unperturbed kernel $\mathbf{K}$ are given as:

$$(\mathbf{K})_{iJ} = \sum_{I \in i} \mathcal{k}(q_J, q_I^*) \tag{3.11}$$

The kernel elements that correspond to the force, i.e. minus the nuclear gradient operator acting on the kernel, are given by:

$$-\frac{\partial}{\partial x_I^*} (\mathbf{K})_{IJ} = -\sum_{K \in i} \frac{\partial \mathcal{k}(q_J, q_K^*)}{\partial x_I^*} \qquad \text{where} \qquad I \in i \tag{3.12}$$

The kernel elements that correspond to the response to the external electric field $\mathbf{E}$ are given by:

$$\frac{\partial}{\partial E_\nu^*} (\mathbf{K})_{i_\nu J} = \sum_{K \in i} \frac{\partial \mathcal{k}(q_J, q_K^*)}{\partial E_\nu^*} \qquad \text{where} \qquad \nu \in \{x, y, z\} \tag{3.13}$$

Similarly, the nuclear Hessian kernel is given by:

$$\frac{\partial^2}{\partial x_{I'}^* \partial x_I^*} (\mathbf{K})_{I'IJ} = \sum_{K \in i} \frac{\partial \mathcal{k}(q_J, q_K^*)}{\partial x_{I'}^* \partial x_I^*} \qquad \text{where} \qquad I', I \in i \tag{3.14}$$

Lastly, the kernel that yields the dipole derivatives necessary for the infrared intensities is written as the mixed second order derivative,

$$\frac{\partial^2}{\partial E^*_\nu \partial x^*_I} (\mathbf{K})_{i_\nu IJ} = \sum_{K \in i} \frac{\partial \mathcal{k} (q_J, q^*_K)}{\partial E^*_\nu \partial x^*_I}$$
$$\text{where} \quad I \in i \quad \text{and} \quad \nu \in \{x, y, z\} \tag{3.15}$$

We are not aware of any other QML model which can account for these effects simultaneously.

### 6.3.3 Comparison to Gaussian Process Regression

In conventional GPR, the response (i.e. derivatives) of the learned function can be included in the training, and the operators are enforced by adding a kernel for each operator of each learned function in the training set.[193] For example, including the nuclear gradient in addition to the energy will add one additional kernel function for each gradient component in the training set. The GPR kernel matrix which simultaneously incorporates the energy, $u$, and the gradient, $g$, is written as:

$$\mathbf{K}^{\text{GPR}} = \begin{bmatrix} \mathbf{K}^{u,u*} & \mathbf{K}^{u,g*} \\ \mathbf{K}^{g,u*} & \mathbf{K}^{g,g*} \end{bmatrix} \tag{3.16}$$

where $\mathbf{K}^{u*,u}$ is the covariance between two molecules, $i$ and $j$. E.g. using a local decompostion this is given by the following double sum:

$$\mathbf{K}^{u,u*}_{ij} = \sum_{I \in i} \sum_{J \in j} \mathcal{k} (q_J, q^*_I) \tag{3.17}$$

Likewise, the first of the two blocks that contain only one derivative are given by

$$\mathbf{K}^{u,g*}_{iKj} = \sum_{I \in i} \sum_{J \in j} \frac{\partial \mathcal{k} (q_J, q^*_I)}{\partial x^*_K} \tag{3.18}$$

and the second block is equal to the transpose. The last block which comprises the largest part of the full kernel matrix is the double derivative given by:

$$\mathbf{K}^{g,g*}_{iKjL} = \sum_{I \in i} \sum_{J \in j} \frac{\partial \mathcal{k} (q_J, q^*_I)}{\partial x_L \partial x^*_K} \tag{3.19}$$

Thus, the memory requirement for a kernel for a training set with $N$ molecules, each with $M$ atoms is dominated by the 2nd derivative covariance kernel which scales as $\mathcal{O}(9N^2M^2)$. With numerical derivatives a gradient is twice as expensive as the kernel itself, and the 2nd derivative is four times as expensive. With these factors, the number of kernel evaulations of the 2nd derivative kernel scales as $\mathcal{O}(36N^2M^4)$.

Within the OQML formalism, as outlined in Sections II A and II B, we do not extend the basis by adding additional kernels functions, but we rather enforce the derivatives of the kernel elements in the regression.

Note that OQML assigns only one $\alpha$ coefficient per atom, regardless the dimensionality of the perturbation. This choice of basis has similarities to the sparsification introduced by Bartók and Csányi,[164] although the mathematical origins are different.

In practice this means that the number of kernel function evaluations needed to train the model is reduced drastically.

The size of the kernel necessary to train our OQML model is Eq. 3.6 is $\mathcal{O}\left(N^2M^2\right)$, regardless of the perturbation. The number of kernel evaluations when the gradient is included for the gradient will scale as roughly $\mathcal{O}\left(6N^2M^3\right)$. For the examples in this work, memory requirements and training times are reduced by factors of $\sim$10 and $\sim$100, respectively, compared to conventional GPR with the same amount of training data.

In GPR the training error will usually be close to zero, since each additional label in the training set will be described by an additional basis kernel function. Since Eq. 3.6 uses a constant number of basis functions, the normal equation will describe an overdetermined set of equations, when the size of the perturbation exceeds the number of basis functions. For example, there are always more gradient components than the number of atoms in a molecule, while for molecules $> 3$ atoms there are always more atoms than dipole moment components. The fact that the problem can become noticeable also means that training errors can become noticeable. Here, we found that in some cases they can even become as large as the test set error.

### 6.3.4 Representation

In this work we extend the Faber-Christensen-Huang-Lilienfeld (FCHL) representation[19] to explicitly include the dependence on an externally applied electric field. This is crucial in order to learn dipole moments and other electric field-dependent properties. The FCHL representation consists of a set of $M$-body expansions $\mathcal{A}_M(I) = \{A_1(I), A_2(I), A_3(I), \ldots, A_M(I)\}$. The terms in the many-body expansion correspond to element type, interatomic distances, and interatomic angles, for the one-, two-, and three-body terms, up to order $M$, respectively.

It has previously been shown that the off-set in the LC is improved when the two- and three-body terms are multiplied by scaling factors such that features that contribute more to the

learned property are weighted higher in the regression.[66] For energy learning, it was shown that $1/r^n$ and an Axilrod-Teller-Muto term[194, 195] are suitable scaling factors for the FCHL two- and three-body terms, respectively.

In this paper, we extend the FCHL representation to include a dependence on the external electric field. Our modified FCHL* representation (denoted by an asterisk) compares the same features as the original formulation (i.e. element type, and interatomic distances and angles), but an extra term is added to the scaling function to emulate the physics of the the electric-field dependence of the representation, and adjust the weighting accordingly. The new two-body scaling function (denoted by an asterisk) is given by

$$\xi_2^{*IJ} = \xi_2^{IJ} - \epsilon(\boldsymbol{\mu}_{IJ} \cdot \mathbf{E}) \tag{3.20}$$

where $\xi_2^{IJ}$ is the $1/r^n$ scaling function in the original FCHL representation, $\mathbf{E}$ is the externally applied electric field, and $\boldsymbol{\mu}_{IJ}$ is a fictitious dipole arising from fictitious partial charges assigned to the atomic site of the atoms $I$ and $J$, and $\epsilon$ is a scaling parameter that balances the two terms in the scaling function. This parameter was fitted *ad hoc* to $\epsilon = 0.005$ Hartree$^{-1}$ using toy models. The center-of-nuclear-charge convention is used to define the origin of the coordinate system. In practice the fictitious partial charges are taken from the Gastieger charge model[196] as implemented in Open Babel.[145] However, we note that the exact values of the fictitious partial charges are unimportant, and any partial charge model could likely be used. Note that the OQML model does not learn these fictitious partial charges, nor does it use these as a proxy to learn the dipole moment. The model learns the scalar field of the energy, and the charges merely serve as dummy variables which enforce the right physical dependence of the kernel elements on the electric field.

The augmented three-body scaling function for an atom $I$ interacting with the atoms $J$ and $K$ is similarly given by:

$$\xi_3^{*IJK} = \xi_3^{IJK} - \epsilon(\boldsymbol{\mu}_{IJK} \cdot \mathbf{E}) \tag{3.21}$$

where $\xi_3^{IJK}$ is the Axilrod-Teller-Muto scaling factor used to weight the three-body terms in the FCHL representation, and $\boldsymbol{\mu}_{IJK}$ is the fictitious dipole arising from fictitious partial charges assigned to the atomic site of the atoms $I$, $J$, and $K$.

In the absence of an externally applied electric field, the FCHL* kernel elements are identical with the original FCHL kernel elements, but the derivative with respect to a perturbing field is now non-zero. We also note that this representation is "non-polarizable"; the second derivative of the representation with respect to the field is zero with a linear kernel. This

could be amended, for example, by using on-site multipoles moments with polarizability tensors, e.g. from a polarizable force field or a chemical-potential equalization charge model, rather than a static charge model.

## 6.4 Results

### 6.4.1 Toy Model for Force Learning

In this section we demonstrate numerically the response of the kernel elements with respect to two very different kinds of perturbations, namely (1) the nuclear coordinates, and (2) an external electric field. The hydrogen fluoride molecule (H-F) is used as a toy model, and to show how including vector quantities in the training improves learning.

We now show how the derivative of the kernel improves learning the potential energy of H-F. The MP2/aug-cc-pVTZ potential energy curve for the H-F molecule is used as training data. Selecting four training points (see Fig. 4.1), models were trained on these four points with and without the interatomic forces in the training set. Not training on forces using the FCHL representation with the default hyperparameters,[19] the resulting model describes the dissociation curve poorly; at the minimum-energy distances it even predicts a spurious transition state, and the energy decreases sharply for $d \to 0$. When the forces are included, however, the potential energy surface is reproduced remarkably almost quantitatively, despite only four points being used to fit the model.

### 6.4.2 Toy Model for Electric Field-Dependent Properties

Here we demonstrate the effect of including the dipole moment in addition to the energy in the training data. We now use a GPR model since our approach in section 6.3.1 would only contain two basis functions, while we are including up to four components, i.e. energy and dipole moment components. The toy model demonstrates the properties of the FCHL* representations which are fully transferable to the ML approach we present herein. We place a H-F molecule in an electric field of 0.001 a.u. which is rotated 360 degrees, and the energy and dipole moment are calculated at each step of 1 degree at the MP2/aug-cc-pVTZ level of theory. We select just one point as training set, and train two GPR models, one with the MP2 energy and dipole moment components and the other with the MP2 energy but without the dipole moment. The energy predictions of these models as a function of the rotation of the field are displayed in Fig. 4.2. Without fitting to the dipole moment, the energy change due to the electric field is

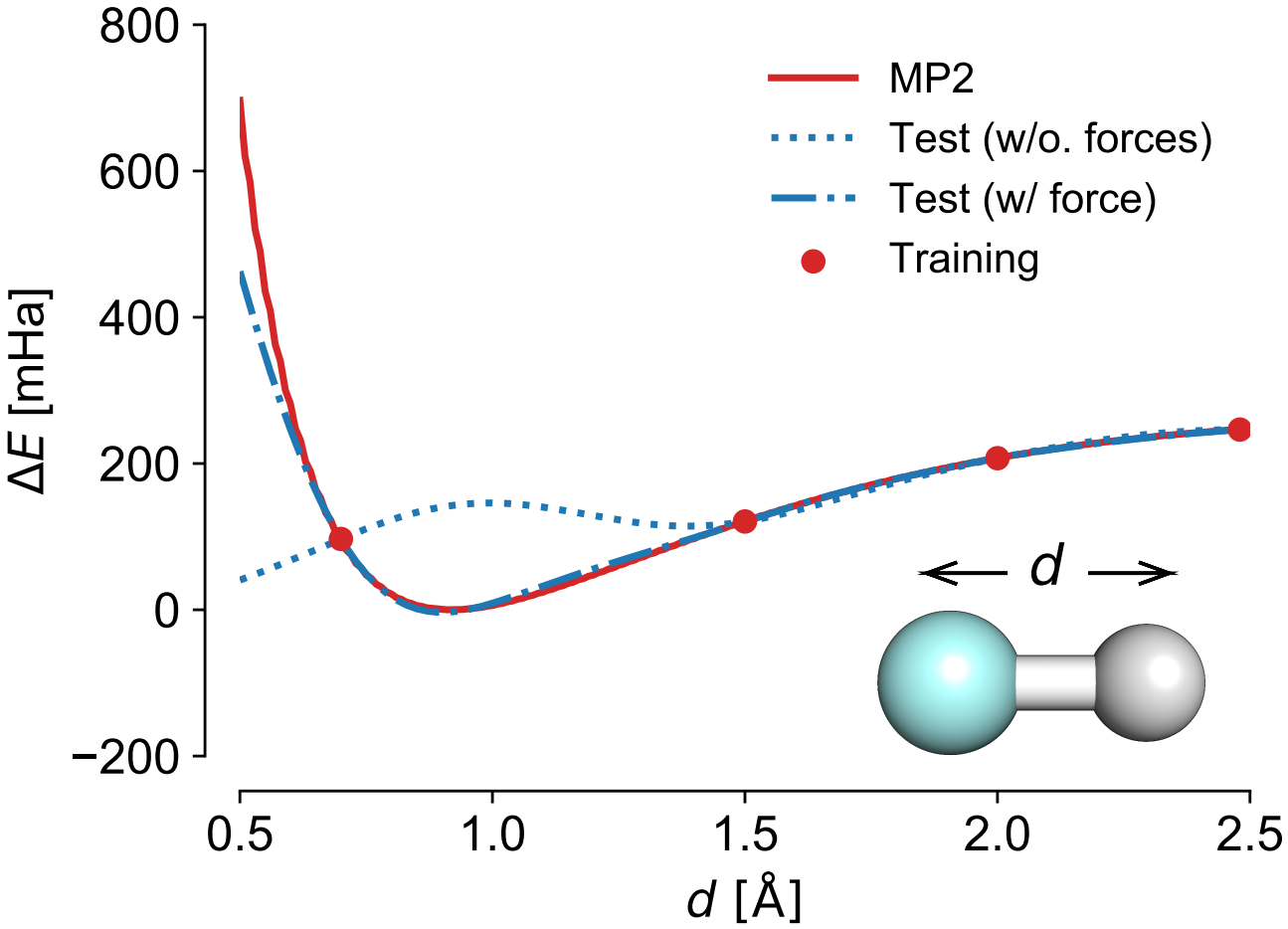



**Figure 4.1:** The MP2/aug-cc-pVTZ potential energy surface of the hydrogen fluoride (H-F) molecule is displayed as a solid red line. Four training points (red dots) are selected and two models are trained and used to predict the potential energy surface: One including the interatomic force in addition to the MP2 energy (blue, dash-dotted), and one using only the MP2 energy (blue, dotted).

close to 0, only fluctuating by a bit of numerical noise from the fit. When the dipole moment is included, the curve is reproduced almost quantitatively with only a negligible deviation at the lowest energy point, presumably due to very small polarization effects and numerical noise. This demonstrates how including a dipole-like dependence on the electric field in the representation is an efficient way to capture the underlying physics of the dipole moment into the kernel.

### 6.4.3 Force and Energy Learning

Here we use the FCHL* representation within the presented OQML model to study two existing benchmark sets for learning forces and energies. The MD17 consists of molecular dynamics (MD) snapshots from MD trajectories of different molecules for which reference forces and energies are available.[61] We are benchmarking our models to seven molecules out of the

MD17 dataset, namely ethanol, salicylic acid, aspirin, malonaldehyde, toluene, naphthalene, and uracil. Similarly, the ISO17 consists of MD snapshots of isomers with the chemical formula $C_7O_2H_{10}$. The ISO17 additionally comes with two different test sets.[38, 118] One that consists only of isomers with a connectivity that is present in the training set ("known") and one that only contains isomers with a connectivity that is not present in the training set ("unknown"). Briefly the two datasets benchmark the conformational freedoms and constitutional freedoms of molecules, respectively. Since there is no electric field applied to the molecules in these data sets, note that the FCHL* representation reduces to the original FCHL representation.[19]

LCs for the two datasets are displayed in Figures 4.3 and 4.4. For reference we compare FCHL* to the Gradient-Domain Machine Learning (GDML) method[61] which is closely related to GPR regression with the inverse distance matrix as representation, and the SchNet neural network.[38] We note that a promising modification to GDML exists, sGDML, which shows higher accuracy compared to GDML for molecules that have atoms that are related by symmetry operations.[197] For the MD17 dataset, the out-of-sample MAE errors of predicted energies are similar between FCHL*, GDML and SchNet, with SchNet being slightly less accurate in most cases (See Fig. 4.3). FCHL* and SchNet perform best for ethanol and malonhaldehyde, while GDML is the best for salicyclic acid and naphthalene. Uracil is best modeled by GDML, with relatively poor SchNet forces, and FCHL being in between. At this point, we remind the reader that the GDML approach is only applicable to a given system, while FCHL* and SchNet are capable of learning across chemical space. Note however, that a direct comparison between the different ML approaches is not possible. Ultimately, the OQML approach is different from SchNet and GDML, not only because of the use of operators, but also in the choice of representation.

Performance across constituational space is tested on the constitutional isomers in the ISO17 dataset (Fig. 4.4). For the two test sets of "known" and "unknown" molecules in the ISO17, the FCHL* model displays a good learning rate, that is qualitatively comparable to the SchNet model. Note that here, the name "known" only implies that the isomers of the same constitution are known to the machine, but not the conformations in the test set. Unfortunately the LCs between the FCHL* models and SchNet do not overlap, so the two models cannot be compared quantitatively here, but the out-of-sample accuracy seems comparable.

Overall, we find that our operator approach leads to forces with state-of-the-art accuracy, on par with two of the most accurate models already published in literature.

### 6.4.4 Learning Dipole Moments of QM9

Prediction errors of machine learning models of dipole moments converge slowly for conventional QML models.[18, 19, 83] Here we demonstrate how including the underlying physics for the dipole moment into the representation improves the learning rate, as opposed to learning the dipole norm with conventional kernel ridge regression. We compare two approaches to learn the dipole moment norm of the molecules in QM9, (1) using the FCHL* representation with the OQML approach outlined in Sec. 6.3.1 to fit the dipole moments as derivatives of the energy and (2) learning the dipole moment norm as a scalar using kernel ridge regression with the FCHL representation as done in our earlier paper.[19] The LCs of the two models are displayed in Fig. 4.5. The MAE out-of-sample predicted dipole moment norm is decreased substantially with our new approach. For instance, training on 5000 random molecules, the out-of-sample MAE error is reduced by 54% (From 0.67 Debye to 0.31 Debye). We also note that not only is the LC offset lower when the dipole moment operator is used, compared to conventional KRR, but it is also substantially steeper. This demostrates the strength of the approach of using the correct response operators in the kernel to learn the corresponding response properties.

### 6.4.5 Learning Normal Modes

In this section we assess the ability of the methodology to predict vibrational normal modes of a number of organic molecules.

We randomly selected 83 molecules from the QM9 dataset with 9 heavy atoms. For each of these molecules we create a minimal training set consisting of all sub-fragments of the molecules with up to 7 heavy atoms, following the methodology of Huang and Lilienfeld.[66] Effectively this approach can be used to prove that the machine can extrapolate from known properties of smaller molecules to predict the same properties for larger molecules.

For each of the these generated fragments, a conformational search is performed using RDKit,[198] and the unique conformers are minimized at the $\omega$B97xD/6-31G(d) level of theory. From each of these minimized geometries, a number of distorted geometries are generated using normal-mode sampling[199] at the same level of theory. For each of the distorted geometries, a single-point energy and force evaluation is performed at the $\omega$B97xD/6-31G(d) level of theory, and the forces and energies are saved. Using the sets of distorted fragment geometries for each of the 83 molecules, we train machines on forces and energies with increasing numbers of samples of each fragment in the sets.

In order to benchmark the performance of the trained machines we set up the following test; a vibrational analysis is performed at the $\omega$B97xD/6-31G(d) level of theory for each of the 83 molecules. Using the normal modes of the molecules obtained from the vibrational analysis, we generate scans of the potential energy surface along each normal mode. The scan consists of structures that are distorted from the equilibrium geometry along each of the normal modes in 10 steps along the positive and negative directions. The distortions along each normal mode are scaled using the force constants, such that the energy of the geometry with the largest distortion along a normal mode is about 0.5 kcal/mol higher than the equilibrium geometry. For each of these potential energy scans along the normal modes, we let the trained machines predict the potential energy, and then we compare this to the QM energy. If the machine predicts a well-defined minimum within the 0.5 kcal/mol scan range, this is counted as a success, otherwise this is counted as a failure. As an example we show predicted normal mode scans for the 15 normal modes with lowest frequency for a QM9 molecule ($C_6N_3H_7$, ID# 036682, SMILES string: `C1C2C3C4OC0C13C24`) in Fig. 4.8. The molecular structure and its corresponding atom-in-molecule fragments (am-ons) used for training can be seen in Fig. 4.6.

In addition, we present predictions from machines trained on $N \in \{1, 2, 4, 8, 16, 32\}$ distorted samples of each sub-fragment in the database. Data to reconstruct similar plots for all 83 molecules is available from Figshare at `dx.doi.org/10.6084/m9.figshare.6994445`. For the machine trained on only $N = 1$ sample per fragment, a total of 11 normal modes do not have a well-defined minimum within the scan range. By increasing the training set to $N = 2$, the machine only predicts two normal modes with minimums outside the scan range. At $N = 4$, all normal modes have a well-defined minimum inside the scan range, but when increasing to $N = 8$, two of the low normal modes that correspond to very non-local conformational changes are not identified correctly to lie within the scan range. Increasing again to $N = 16$ samples, the minimums are well-defined again, and at $N = 32$, the QM potential energy curves are almost quantitatively reproduced.

We note that the higher normal modes, which mostly correspond to very local distortions such as a single hydrogen bond stretching, are almost always very well reproduced. In contrast, the lower normal modes, which often are more non-local in nature and correspond to very flat energy surfaces, require larger training set sizes to reproduce correctly.

Repeating the same test for all of the 83 QM9 molecules, we can plot the fraction of normal modes which are incorrectly described as function of the training set size. Here, training set size is measured as the maximum possible rank of the kernel matrix, which corresponds to the

number of regression coefficients and the number of atoms in the training set. This is plotted for all 83 molecules in Fig. 4.7 for the corresponding machines training on $N \in \{1, 2, 4, 8, 16, 32\}$ distorted samples of each sub-fragment. We note a trend that larger training sizes yield a smaller chance that the machine fails to identify a well-defined minimum close to the minimum in the reference geometry.

### 6.4.6 Infrared Spectrum for Dichloromethane

In order to demonstrate the utility of the above developments, we have combined them in order to learn and predict IR spectra. More specifically, a vibrational analysis is performed to get the harmonic frequencies and the IR intensities for the dichloromethane molecule. We note that although our methodology is transferable, the results of this exercise is very dependent on the training set. Thus we restrict this section to only one molecule, and demonstrate that the methodology yields higher order derivatives, including mixed derivatives that systematically improve with the training set.

Models are trained on distorted geometries of the dicholoromethane molecule for which MP2/def2-TZVP energies, forces, and dipole moments had been calculated previously. The training set consists of 100 distorted geometries which are generated by normal-mode sampling following the protocol of Smith *et al.*[199] Using the trained model, a standard vibrational analysis using the rigid-rotor harmonic-oscillator approximation is performed in a standard quantum chemistry package (Gaussian09)[200] via an interface to the QML code[201] which supplies the necessary energies and derivatives to the quantum chemistry program. First, the molecule is optimized on the machine learned potential energy surface by supplying the optimizer in the Gaussian program with the energies and nuclear gradients. Secondly, the vibrational analysis is performed by supplying the Gaussian program with the numerical nuclear Hessian and dipole derivatives. As a reference we compare the IR spectrum from the vibrational analysis on the potential energy surface of the machine learning model to the IR spectrum from a standard vibrational analysis at the MP2/def2-TZVP level.

Five models are trained on a decreasing number of samples (100, 50, 25, 10, 5) of randomly selected configurations from the full 100 configurations training set. Then, a geometry optimization and a vibrational analysis is performed with each of the trained models. The resulting IR spectra for dichloromethane are displayed in Fig. 4.9. Qualitatively the FCHL* models reproduce the frequencies of the true MP2 reference with close agreement between the vibrational frequencies of the tallest peaks, even with as few as 10 training samples. In the spectrum gener-

ated using the largest training set (100 samples), the three most intense peaks in the spectrum are located at 743, 793 and 1318 $cm^{-1}$, compared to 740, 793 and 1315 $cm^{-1}$ for the reference MP2 spectrum. Training the model on only five randomly selected samples does not lead to a meaningful IR spectrum; however, already with ten instances, decent frequencies and underestimated intensities are obtained for the first two peaks. Learning the intensities via the dipole derivatives seem to be a harder task for the machine, compared to the peak locations, and the relative peak intensities are not qualitatively correct until $N = 50$ training samples.

We note that the dichloromethane molecule has 9 normal modes, and it is therefore expected that at the very least 9 samples would be necessary to have the minimally required sampling along all the possible normal modes. Further increasing the training set size to 25 and 50 samples improves the locations of the peaks to MAE vibrational frequencies of 25.6 and 5.7 $cm^{-1}$, respectively. At 100 training samples the spectrum is almost at spectroscopic precision with an MAE of only 2.5 $cm^{-1}$.

This demonstrates the generality of the response operator-based machine learning model. The IR intensities correspond to a second order mixed derivative, indicating that the model accounts even for higher order effects after including only energy and first order derivatives. These results suggest that the systematic addition of higher order effects has the potential to improve the performance even further.

## 6.5 Methodology

### 6.5.1 Used Software

All energy, gradient, and dipole-moment calculations for the H-F molecule were performed in ORCA 4.0.1[203] at the MP2/aug-cc-pVTZ level of theory with no RI approximation and the `NoFrozenCore` keyword. The relaxed MP2 density was used to calculate the dipole moment as the correct derivative of the energy.

Since only the dipole norms are supplied with the QM9 dataset,[63, 86] the dipole moment vectors of QM9 were re-calculated using ORCA 4.0.1. To ensure consistency with the B3LYP/6-31G(2df,p) method and basis set used in the original QM9 dataset, the `B3LYP/G` option was used for the B3LYP functional[90] and the 6-31G(2df,p) basis set was manually set up to the same contraction coefficients and exponents as used in the original calculations.

Energies, forces, and vibrational analyses for the QM9 molecules and fragments in section 6.4.5 were calculated at the $\omega$B97xD/6-31G(d) level of theory using the Gaussian09 program.[200]

The structures and corresponding data can be found in comma-separated values format from Figshare at `dx.doi.org/10.6084/m9.figshare.7000280`.

The forces, energies and dipole moments of the dichloromethane molecule were calculated at MP2/def2-TZVP level of theory in the Gaussian09 program. The MP2 vibrational analysis was also carried out in Gaussian09. The vibrational analyses that employ machine learning were also carried in Gaussian09 via a Python interface to the machine learning code, and the keywords `freq=(numer,fourpoint,step=100)` was used to get the second derivatives. Our current implementation employs two-point numerical first derivatives, except for geometry optimizations for which it was necessary to use a five-point numerical derivative due to the sensitivity to numerical noise in the optimizer.

The reader can carry out machine learning with the presented algorithms, i.e. implemented kernel functions, efficient solvers and the FCHL* representation. The necessary code is freely available from our open source machine learning toolkit QML[201] at `http://github.com/qmlcode/qml`.

### 6.5.2 Hyperparameters

All hyper parameters of the FCHL* representation were kept fixed to the same values as those found to be optimal in our previous paper,[19] and the only new parameter is the newly introduced $\epsilon = 0.0005$ Hartree$^{-1}$ parameter in the scaling functions. In all examples, a Gaussian kernel function is used with the kernel width set to $\sigma = 0.64$, and the cap for smallest singular values to keep in the SVD decomposition was set to $10^{-9}$ in units of the largest singular value. These parameters were not rigorously fitted to any dataset, so it is possible that more optimal values exist.

## 6.6 Conclusion

This paper explores a kernel-based supervised machine learning model that is capable of learning response properties by applying the corresponding response operator to the kernel function. Within this framework, we have extended the FCHL representation by a physically motivated response term for the application of an external electric field. Using the hydrogen fluoride molecule as a toy model, we have demonstrated how the machine learning model and representation can account for the right physics in simple systems with only a minimal number of training samples. Benchmarking the accuracy of our model for force and energy prediction on

the MD17 and ISO17 dataset, our OQML models achieve state-of-the-art accuracy, on par or better than the GDML and SchNet models. For learning the dipole norm of the molecules in the QM9 dataset, using the operator formalism leads to an improvement of 54% compared to learning the same quantity as a scalar with the same representation. Lastly we allude to the possibility to obtain higher order derivatives, including mixed derivatives. This idea has been demonstrated by training a model on the energies, forces, and dipole moments for the dicholoromethane molecule. Using the resulting model we have performed a vibrational analysis and presented the resulting infrared spectrum which systematically approaches the reference spectrum (calculated at the corresponding *ab initio* level of theory) as more training cases are being added.

Our results suggest that it is advantageous to learn response properties via the corresponding response operators. The OQML methodology presented here is, in principle, not limited to derivatives of the energy with respect to the nuclear positions or the external electric field. We envision extending the representation to account for a multitude of other properties, such as higher order response properties, including magnetic properties such as NMR chemical shifts and spin-spin coupling constants, or alchemical derivatives. Since the OQML formalism is not restricted to any choice of operator, it might also be possible to go beyond response operators. For instance, with the right representation, it should be possible to even learn more fundamental properties of molecules such as the electronic density or the kinetic energy.

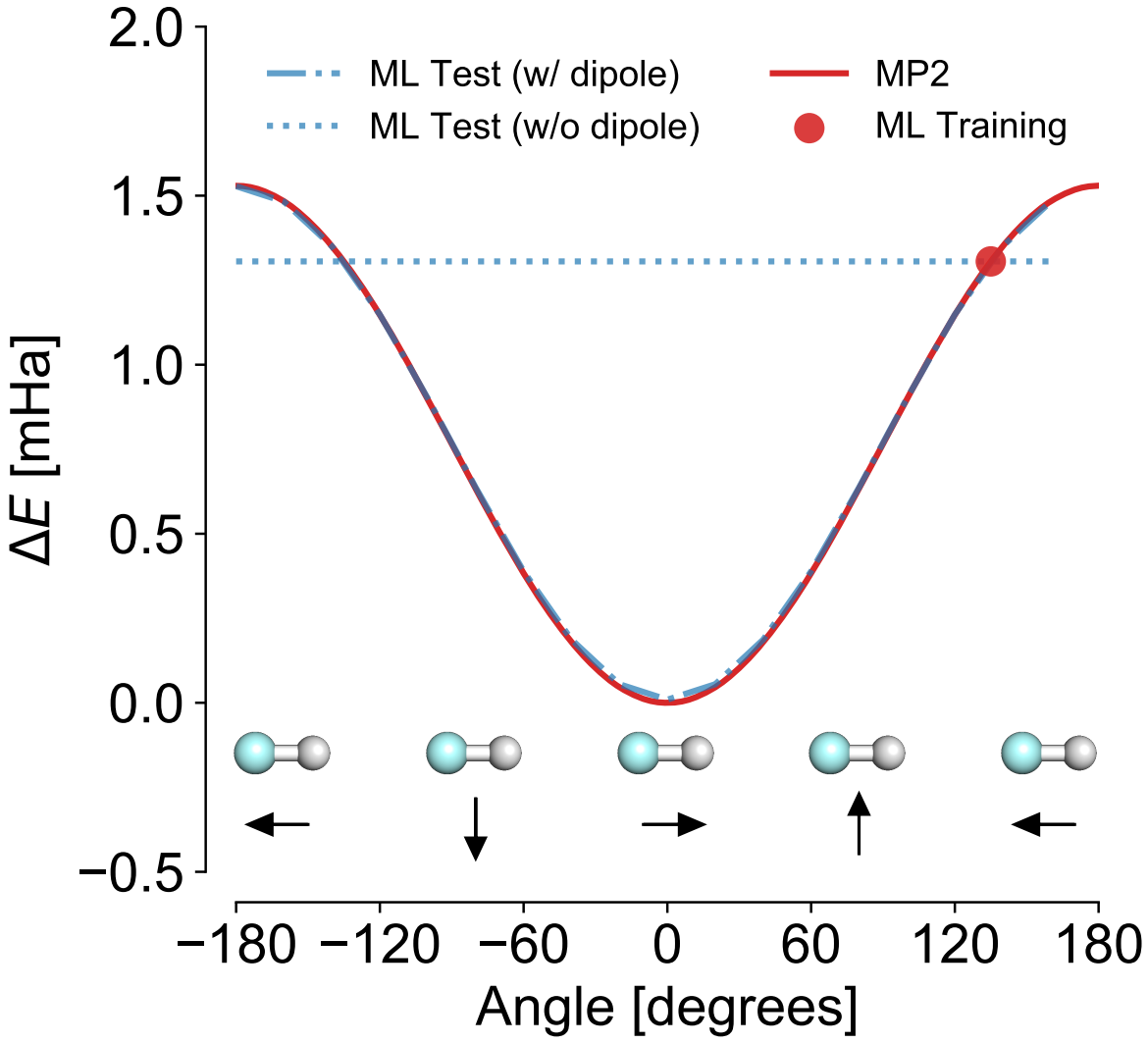



**Figure 4.2:** A hydrogen fluoride (H-F) molecule is placed in an external electric field of 0.001 a.u., and the MP2/aug-cc-pVTZ energy is calculated as a function of the angle between the H-F molecule and the field vector, displayed as a red line. A single point is selected as training set (red dot), and two models are trained and used to predict the energy in the electric field: one including the dipole moment of the molecule in addition to the MP2 energy (blue, dash-dotted), and one using only the MP2 energy (blue, dotted). The alignment between the field and the molecule is sketched at the bottom for clarity.

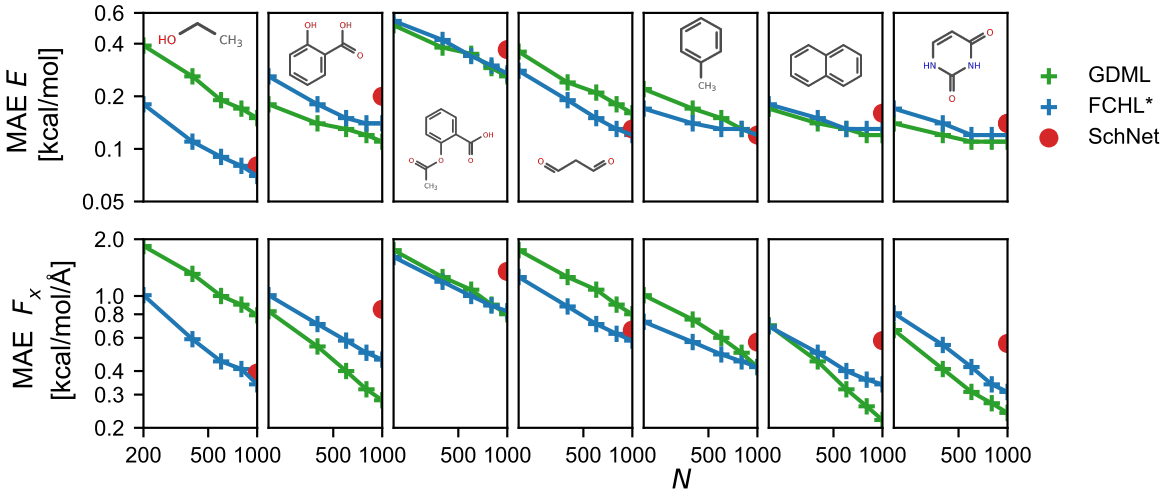



**Figure 4.3:** The two figures show the LCs of our model for the MD17 dataset, for the seevn molecules in the MD17 dataset (from left to right) ethanol, salicylic acid, aspirin, malonaldehyde, toluene, naphthalene, and uracil. The out-of-sample MAE energy prediction ($E$, top row) and MAE force component prediction ($F_X$,bottom row) are shown for the presented FCHL* (blue) model as well as for the GDML[61] (green) and SchNet models (red).[38, 118]

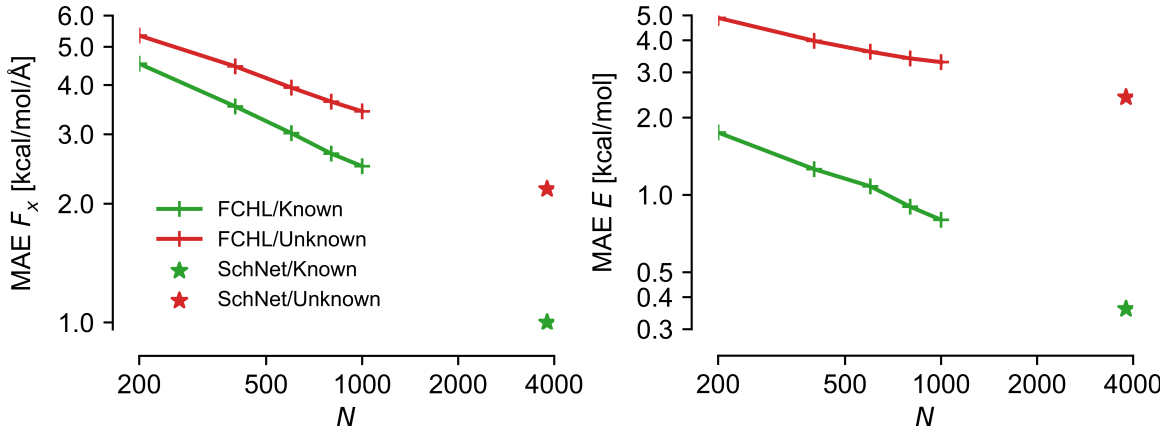



**Figure 4.4:** The two figures show the LCs of our model for the ISO17 dataset, in addition the accuracy for SchNet when using 4,000 training samples is shown. Left shows the out-of-sample MAE energy prediction for a set of isomers known to the trained machine ("known") and for a set of unknown to the machine ("unknown"). Right shows the out-of-sample MAE force prediction for the same two sets. Note that "known" in this context only concerns whether the isomers are included in the training set or not. In both cases only isomers with a conformation unknown to the machine are used to as test data.

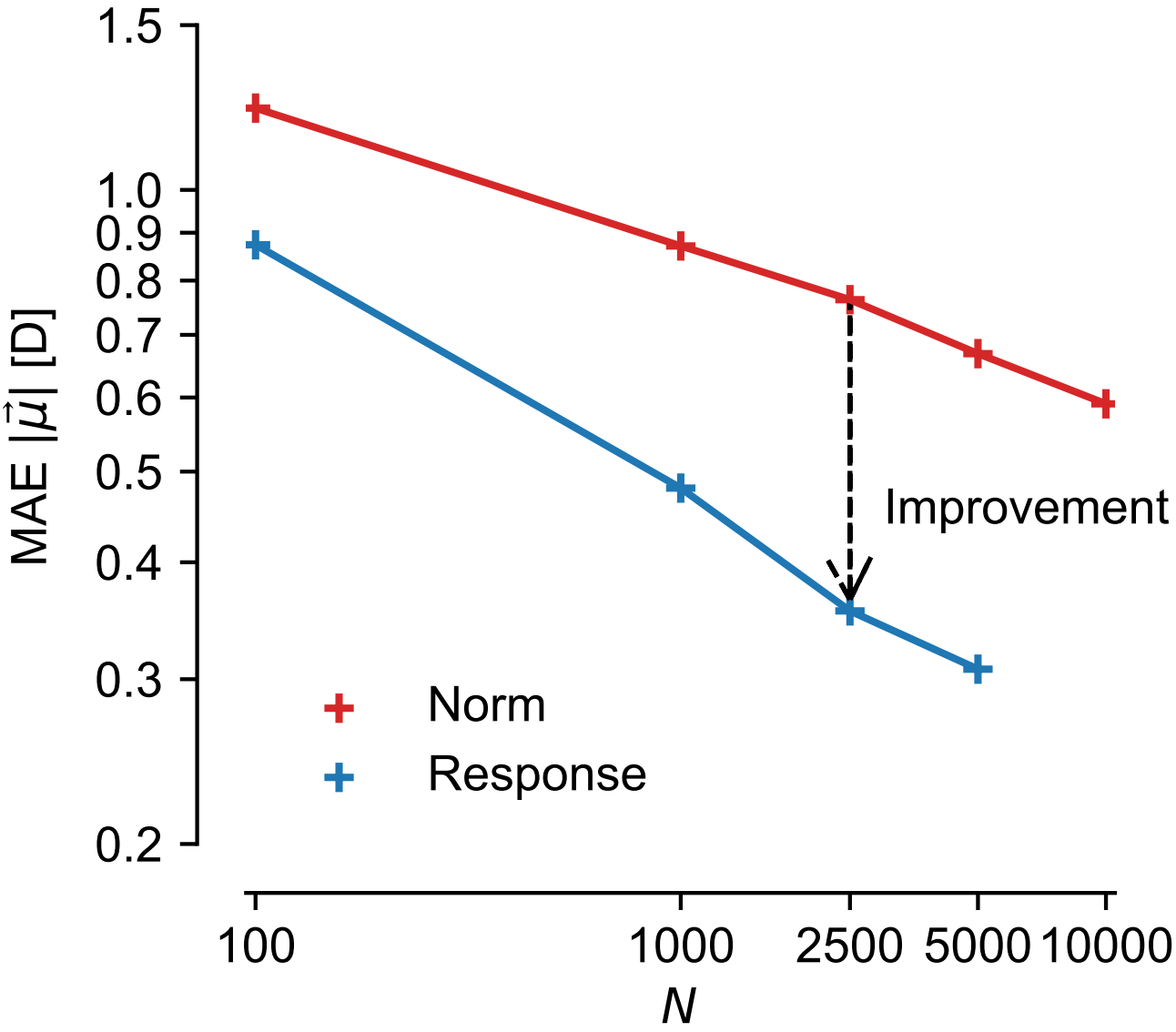



**Figure 4.5:** The figure displays the out-of-sample prediction error of the dipole norm as a function of the QM9 training data set size. The red curve corresponds to a conventional KRR model learning the scalar with the original FCHL representation, taken from Faber *et al.*[19] The blue curve shows the predictions from a machine trained on the energy and dipole moments of QM9 molecules, which in turn predicts the dipole vector from which the norm is calculated.

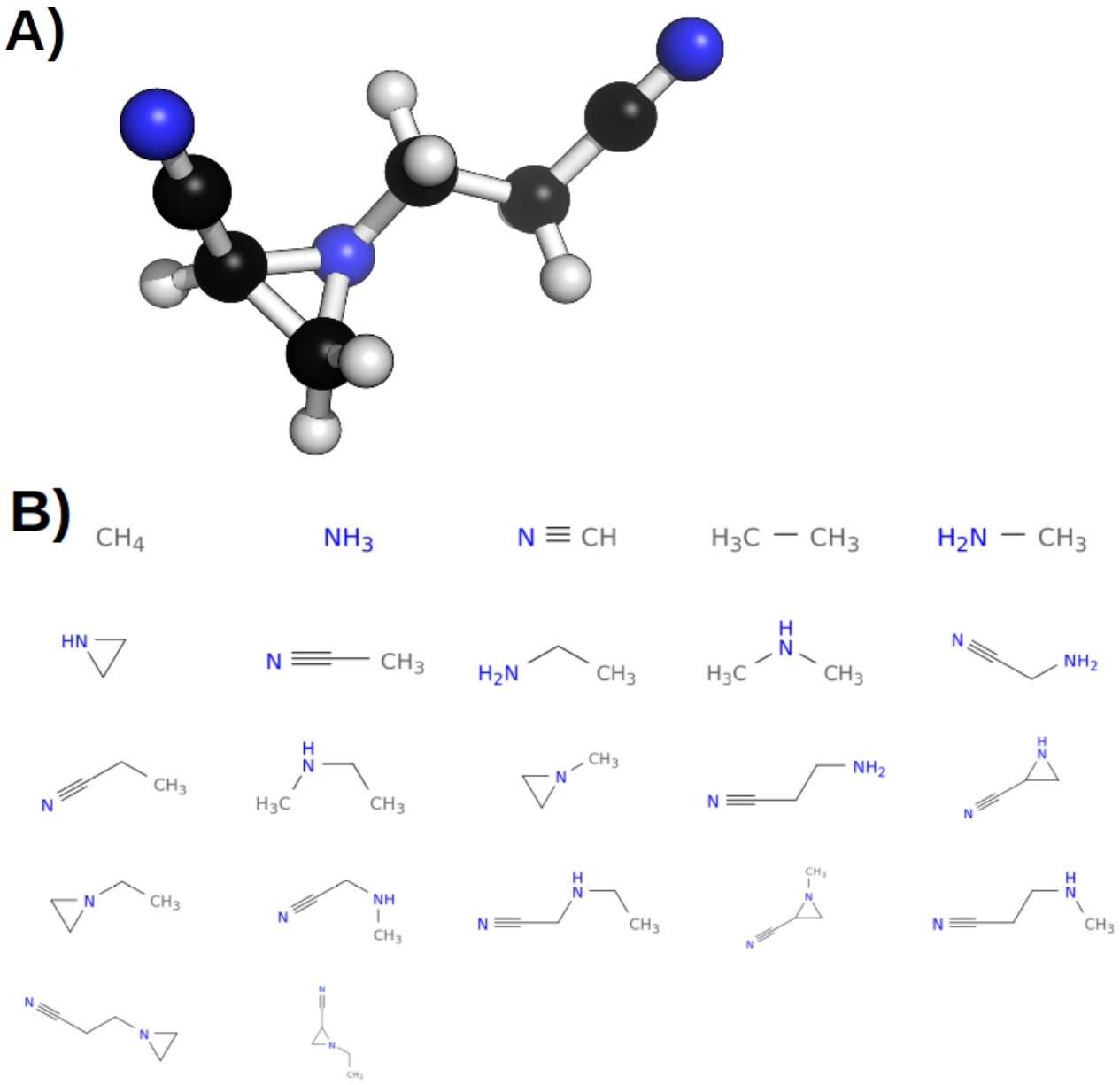



**Figure 4.6:** A) displays the QM9 molecule with the ID# 036682 (SMILES string: `C1C2C3C4OC0C13C24`) for which normal modes have been predicted in Fig. 4.7. B) displays the fragments identified using the method of Huang and Lilienfeld,[66] which are used to generate the training set for the molecule.

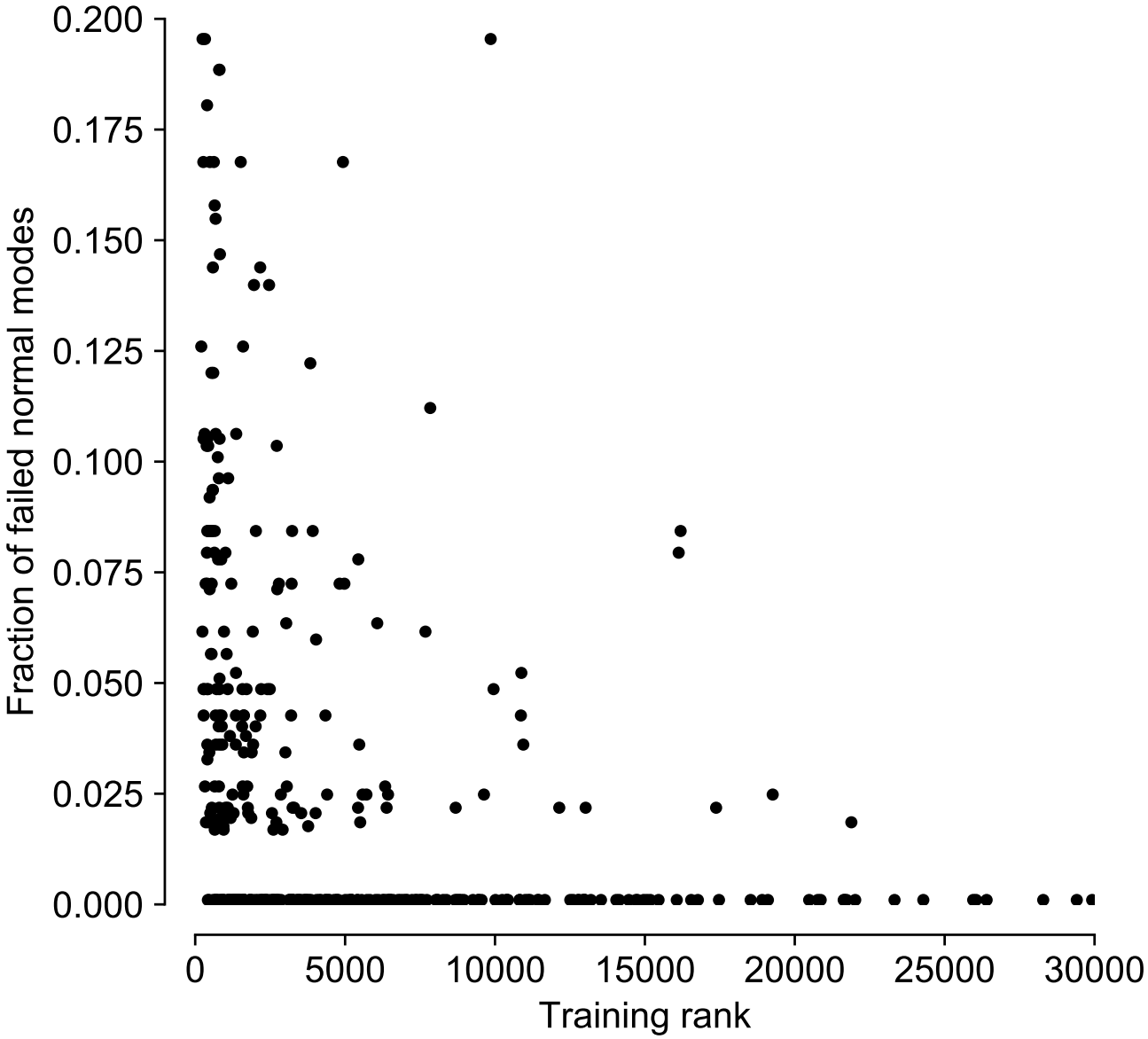



**Figure 4.7:** Fraction of failed normal mode predictions for 83 QM9 molecules with 9 heavy atoms as a function of training set size. For each molecule six machines are trained with increasing numbers of molecules in the training set. The X-axis shows the rank of the kernel matrix (i.e. the number of regression coefficients) for each training set used to train a model for a molecule. The Y-axis shows the fraction of modes for the same molecule, for which the machine predicts a well-defined minimum within a reasonable distance (see text) from the reference equilibrium geometry.

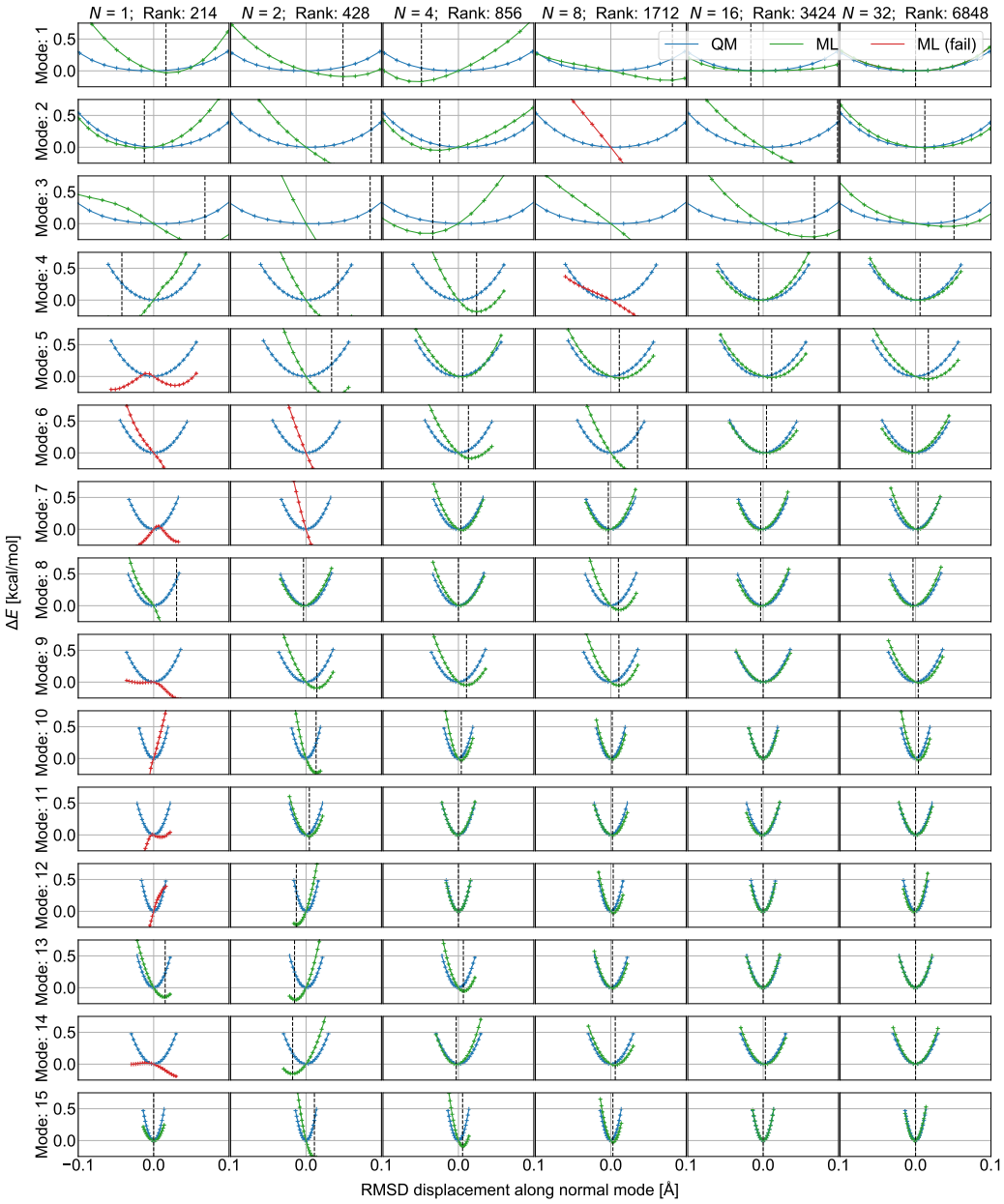



**Figure 4.8:** ML predicted energy changes of $C_6N_3H_7$ as a function of distortion along each of the 15 normal modes with lowest frequency. The molecular structure and its corresponding atom-in-molecule fragments used for training can be seen in Fig. 4.6. Stiffer normal modes are easier to learn and therefore not shown. The complete results set is provided in the SI. Each row and column correspond to a normal mode and training set size N/maximum possible rank of kernel matrix, respectively. N is the number of samples for each amon (i.e. sub fragment). Displacements are scaled such that the maximum distortion energy is close to 0.5 kcal/mol. The X-axis displays the RMSD difference in coordinates to the QM equilibrium geometry after the molecule has been displaced along that normal mode. The Y-axis is the energy difference to the equilibrium geometry, either calculated with QM (blue) or ML (green/red). The curves predicted from ML are displayed in green if there is a defined minimum within the scan range, and red (fail) otherwise. The locations of the minima marked by black, vertical, dashed lines.

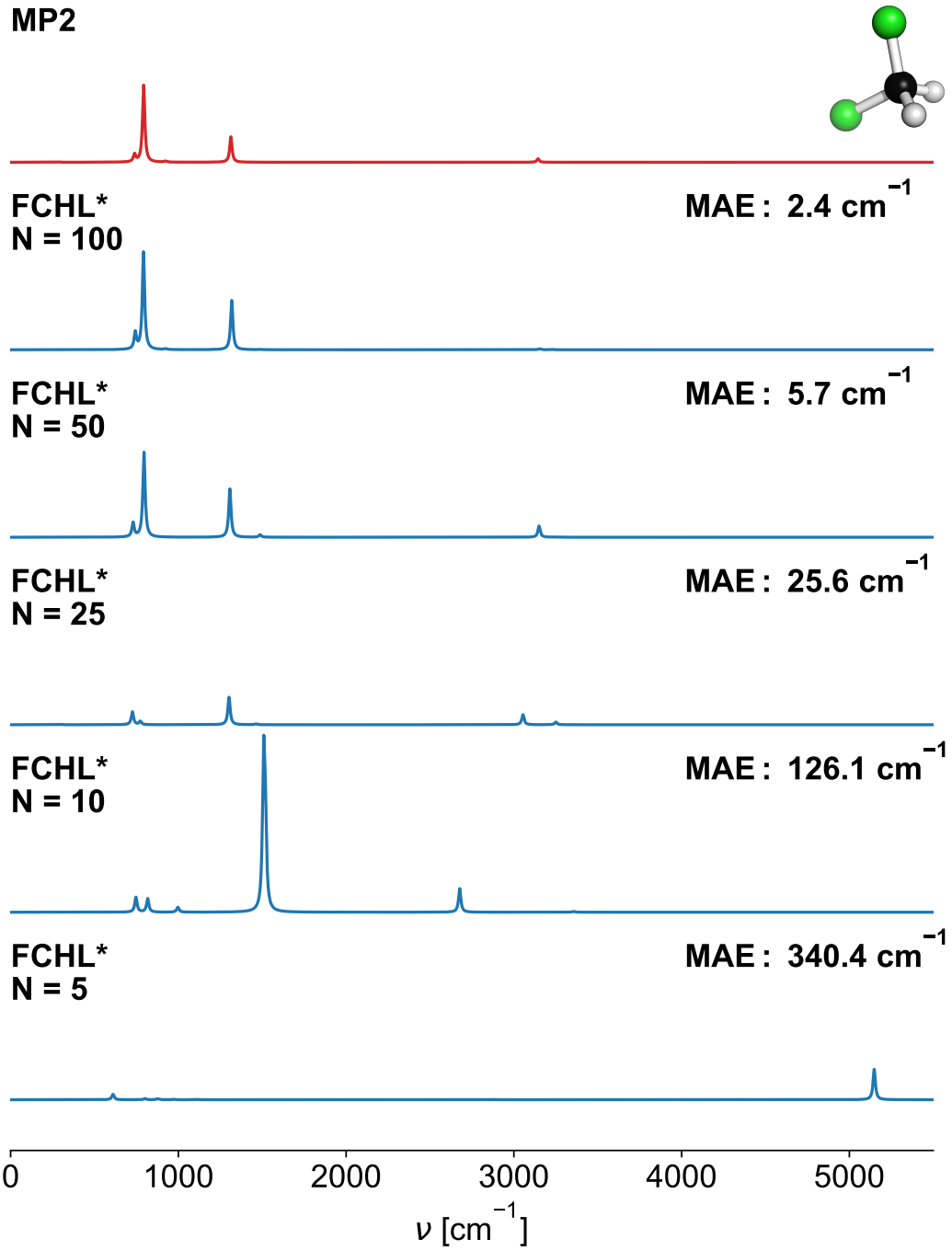



**Figure 4.9:** The unscaled infrared spectrum of dichloromethane calculated via vibrational analysis. (Top/red) calculated at the MP2/def2-TZVP level of theory; (bottom/blue) using QML to calculate the necessary derivatives of the energy with respect the nuclear coordinate and the dipole moment. The spectra are convoluted using Lorentzian distributions[202] with a width of $\gamma = 8$ cm$^{-1}$.

# Chapter 7

# Machine Learning Energies of 2 Million Elpasolite ($ABC_2D_6$) Crystals



## 7.1 Executive Summary

Elpasolite is the predominant quaternary crystal structure ($AlNaK_2F_6$ prototype) reported in the Inorganic Crystal Structure Database. We have developed a machine learning model to calculate density functional theory quality formation energies of all ∼2 M pristine $ABC_2D_6$ Elpasolite crystals which can be made up from main-group elements (up to bismuth). Our model's accuracy can be improved systematically, reaching 0.1 eV/atom for a training set consisting of 10 k crystals. Important bonding trends are revealed, fluoride is best suited to fit the coordination of the D site which lowers the formation energy whereas the opposite is found for carbon. The bonding contribution of elements A and B is very small on average. Low formation energies result from A and B being late elements from group (II), C being a late (I) element, and D being fluoride. Out of 2 M crystals, 90 unique structures are predicted to be on the convex hull—among which $NFAl_2Ca_6$, with peculiar stoichiometry and a negative atomic oxidation state for Al.

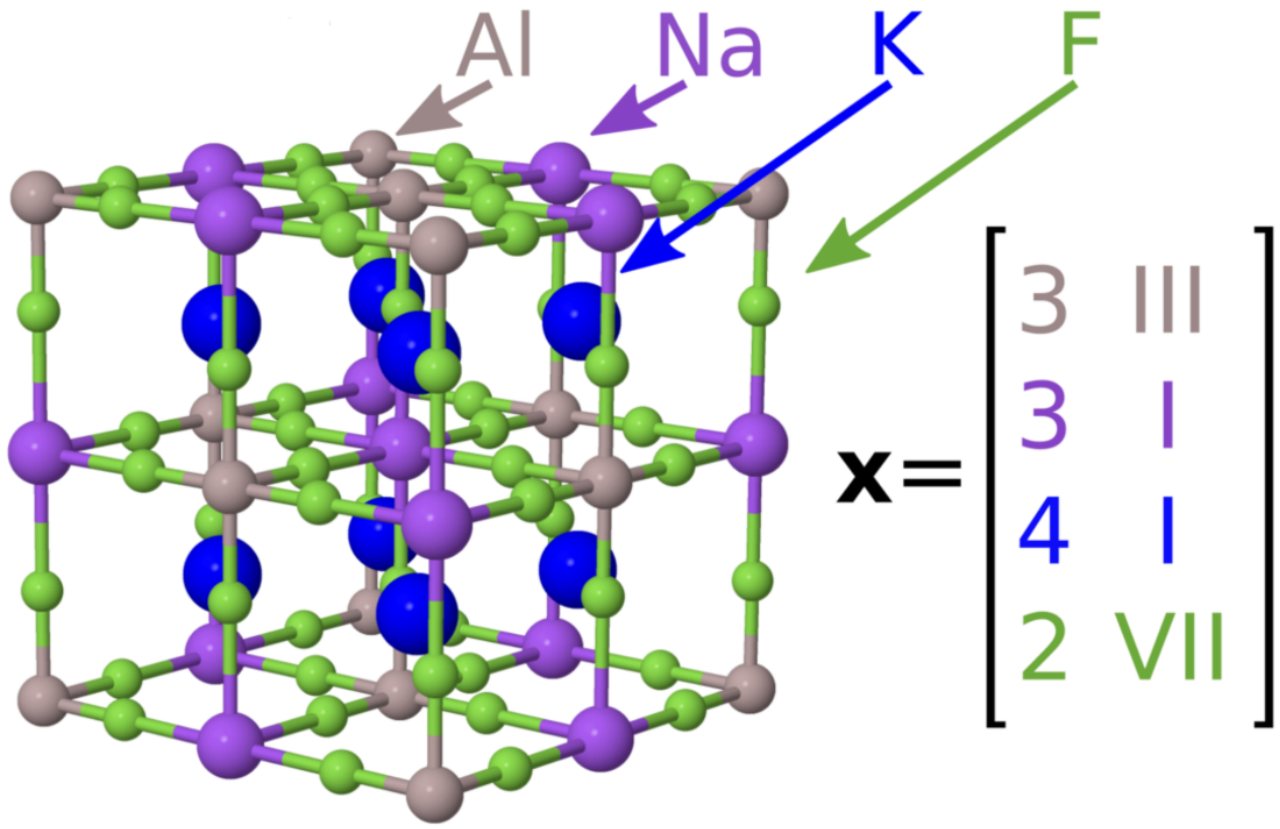



**Figure 2.1:** Illustration of Elpasolite crystal ($AlNaK_2F_6$ structure). The four-tuple $x = (x_1, \ldots, x_4)$ representation of atomic sites is specified.

## 7.2 Introduction

Elpasolite ($AlNaK_2F_6$) is a glassy, transparent, luster, colorless, and soft quaternary crystal in the Fm3m space group which can be found in the Rocky Mountains, Virginia, or the Apennines. The Elpasolite crystal structure (See Fig. 2.1) is not uncommon, it is the most abundant prototype in the Inorganic Crystal Structure Database [106, 107]. Some Elpasolites emit light when exposed to ionic radiation, which makes them interesting material candidates for scintillator devices [204, 205]. One could use first-principle methods such as DFT [9, 10] to computationally predict the existence and basic properties of every Elpasolite. Unfortunately, even when considering crystals composed of only main group elements (columns I to VIII) the sheer number of the $\sim$2 M possible combinations makes DFT based screening challenging—if not prohibitive. Recently, computationally efficient ML models were introduced for predicting molecular properties with the same accuracy as DFT [33, 82]. Requiring only milliseconds per prediction, they represent an attractive alternative when it comes to the combinatorial screening of millions of

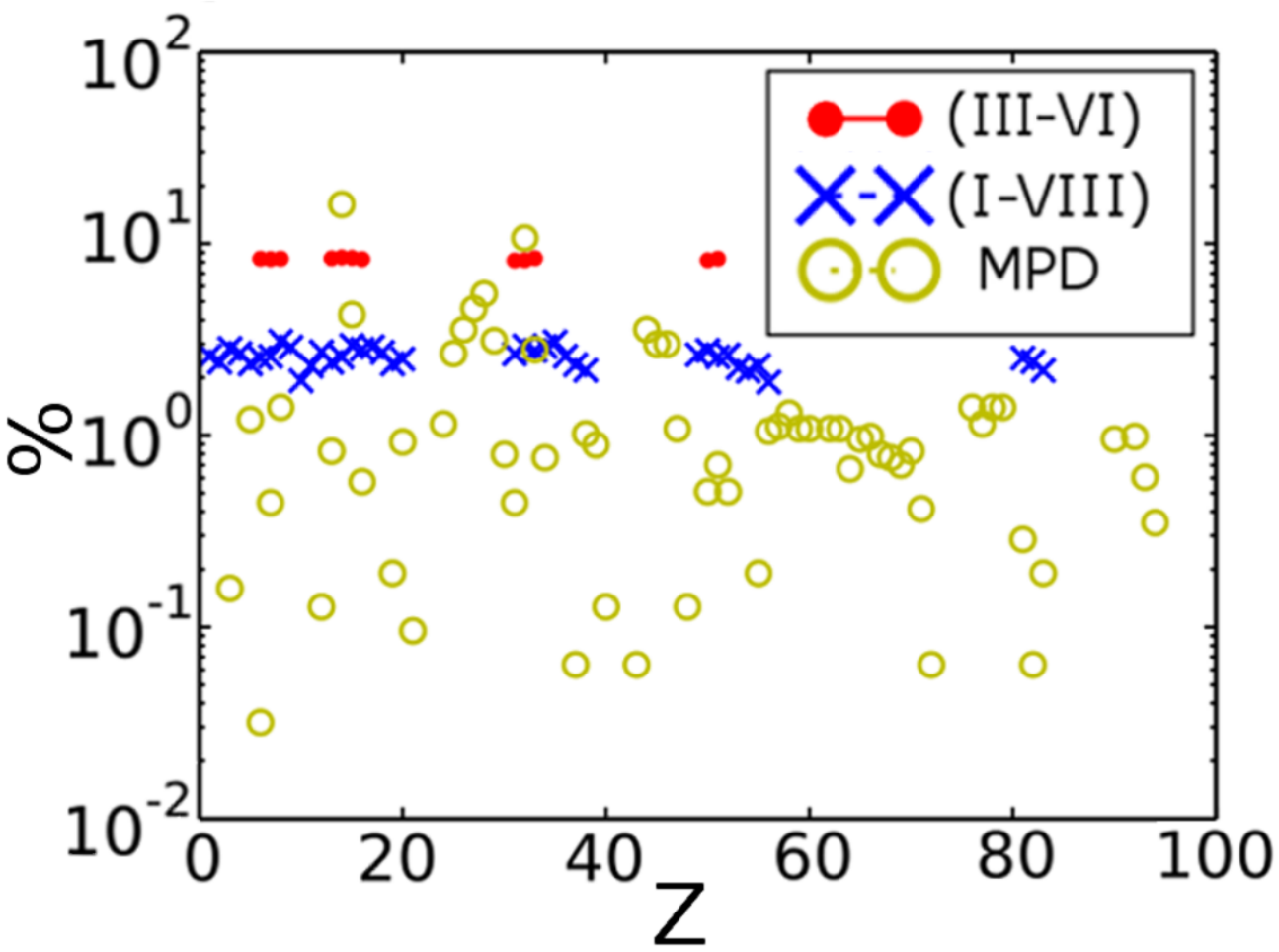



**Figure 2.2:** Frequency of elements (defined by nuclear charge Z) for the three data sets studied.

crystals. While some ML model variants have already been proposed for solids [160, 174, 206], a generally applicable ML-scheme with DFT accuracy of formation energies is still amiss.

In this Letter we introduce a newly developed ML model which we use to investigate the formation energies of *all* ∼2 M Elpasolites made from all main-group elements up to Bi.

Resulting estimates enable the identification of new elemental order of descending Elpasolite formation energy, crystals with peculiar atomic charges, 250 Elpasolites with lowest formation energies, as well as 128 new crystal structures predicted to lie on the convex hull among which $NFAl_2Ca_6$, an Elpasolite with unusual composition and atomic charge.

The ML model achieves the same, or better, accuracy with respect to DFT as DFT in comparison to experimental data and can be generalized to any crystalline material.

## 7.3 Methods

### 7.3.1 Machine Learning Model

The ML-model is based on KRR [22, 23, 25] which maps the non-linear energy difference between the actual DFT energy and an inexpensive approximate baseline model into a linear feature space [31]. More specifically, we construct a ML model of the energy difference between the crystal energy and the sum of static, atom-type dependent, atomic energy contributions $\epsilon_{It}$, obtained through fitting of each atom type $t$ in all main group elements up to Bi. The ML-model is a sum of weighted exponentials in similarity $d$ between query and training crystal. The total energy-predicting model function reads

$$E(\mathbf{x}) = \sum_{I}^{N'} \epsilon_{It} + \sum_{i}^{N} \alpha_i e^{-d_i/\sigma}, \tag{3.1}$$

where $N'$ is the number of atoms/unit cell (10 in the case of Elpasolites), and the second sum runs over all $N$ training instances. $\{\alpha_i\}$ are the weights obtained through linear regression, and $\sigma$ is the global exponential width, regulating the length scale of the problem. The similarity $d_i$ is the Manhattan distance, i.e., $d_i = \|\mathbf{x} - \mathbf{x}_i\|_1$. While various crystal structure representations $\mathbf{x}$ have previously been proposed [160, 174, 176, 206, 207], we have found the following representation to yield superior performance: $\mathbf{x}$ is a $n \times 2$ tuple that encodes any stoichiometry within a given crystal prototype. For quaternary ($n = 4$) Elpasolites, each $x_{1-4}$ refers to the 4 representative sites, the atom type for each site is represented by its row (principal quantum number 2 to 6) and column (number of valence electrons) I to VIII in the periodic table, and sites are ordered according to the Wyckoff sequence of the crystal. As such, $\mathbf{x}$ implicitly represents the energy minimum structure for a system restricted to this prototype—without explicitly encoding precise coordinates, lattice constants, or other (approximate) solutions to Schrödinger's equation. This representation is not restricted to the Elpasolite structure, it can be used for any fixed crystal symmetry: Below we also briefly discuss test results for small size ML models applied to ternary crystals.

### 7.3.2 Data set

For training and evaluation, we have generated DFT formation energies for two data sets of Elpasolites (for computational details see chapter 3), one small, (III−VI), made up from only 12 elements, C, N, O, Al, Si, P, S, Ga, Ge, As, Sn, and Sb; and one large, (I−VIII), containing all main-group elements up to Bi. Since (III−VI) only comprise ∼12 k possible permutations,

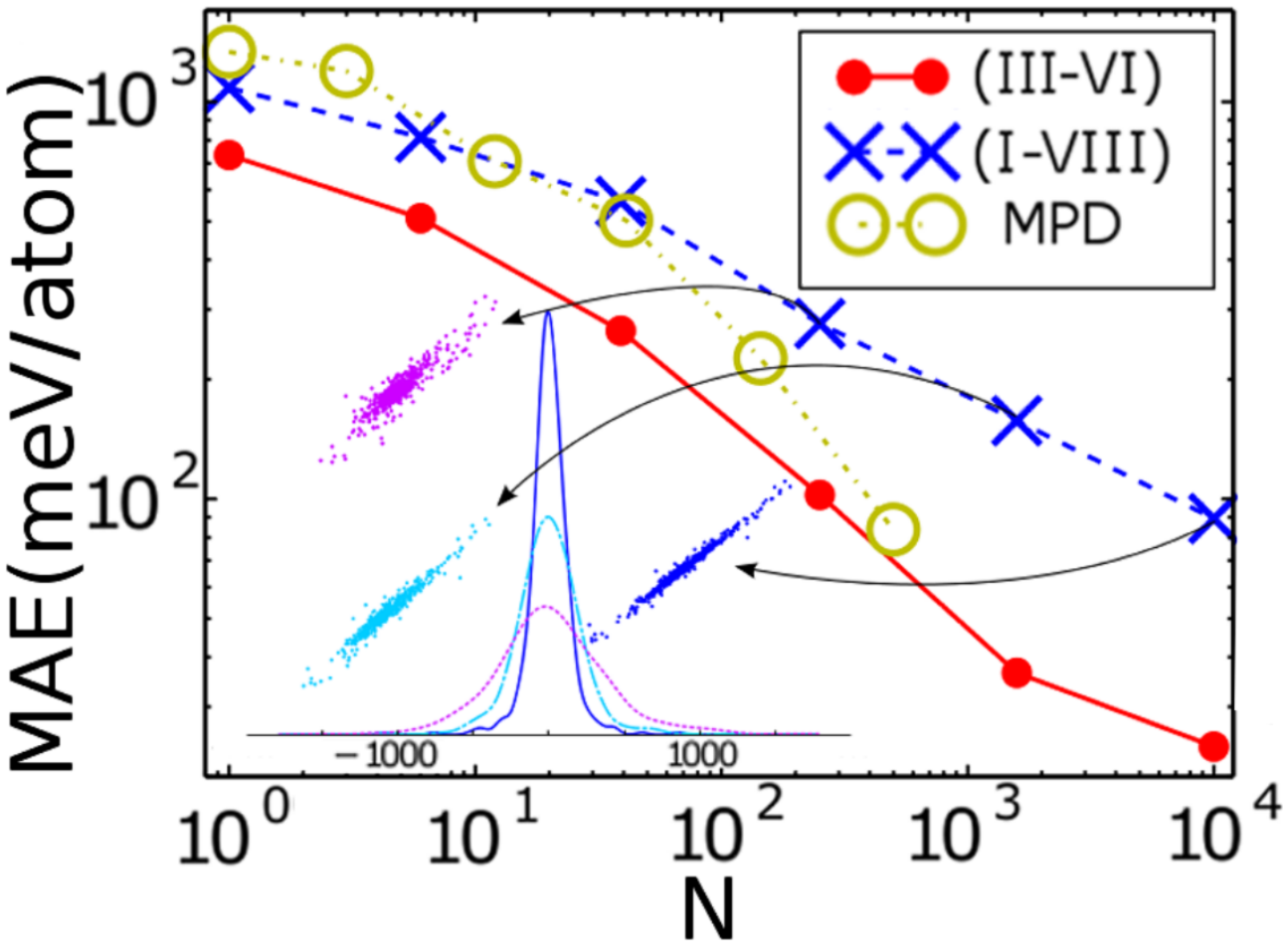



**Figure 3.3:** Mean absolute out-of-sample prediction error as a function of training set size for the three data sets studied. Inset: Error distributions and DFT vs. ML scatter plots for three training set sizes for the (I−VIII) data set.

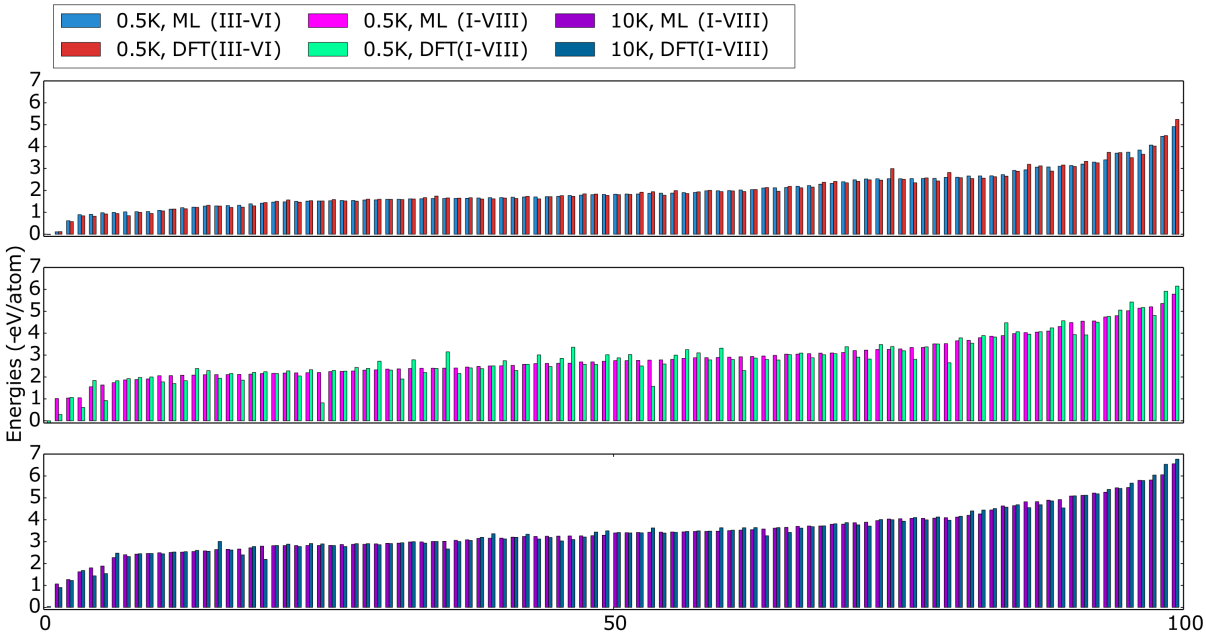



**Figure 4.4:** Calculated ML and DFT formation energies of 100 Elpasolites drawn at random from (III−VI) (TOP) and (I−VIII) (MID & BOTTOM) data sets. Vertical axis corresponds to ML-predicted formation energy (eV/atom) relative to lowest lying structure. TOP and MID ML models have been trained on 500 crystals; the ML model used for the BOTTOM panel has been trained on 10 k crystals. Respective MAEs are 0.303, 0.151, and 0.102 eV/atom.

we have obtained the complete list of formation energies.

## 7.4 Results and discussion

(I−VIII) consists of 10 k structures, i.e. 0.5% of the total number of 2 M possible crystals. The (I−VIII) data set has been generated through random selection of Elpasolites while ensuring an unbiased composition. To verify that the ML model is general and not only restricted to Elpasolites, we have also included a MP [208] dataset (MPD) consisting of ∼0.5k ternary crystals in $ThCr_2Si_2$ (I4/mmm) prototype and made up of 84 different atom types. The distribution of the chemical elements in the data sets are shown in Fig. 2.2.

Numerical results on display in Fig. 3.3 indicate systematic improvement of the predictive accuracy of the ML model with increasing training set size, for all three datasets. The inset details normally distributed errors and scatter plots which systematically improve with training set size for the models trained on the (I−VIII) datasets. For a 10k training set, the ML model reaches a MAE of 0.1 eV/atom compared to reference, i.e. semi-local DFT. DFT, in turn, has an estimated MAE of ∼0.19 eV/atom compared to experiments on heats of formation for general chemistries with filled *d*-shells [209]. For transition metal oxides and elemental

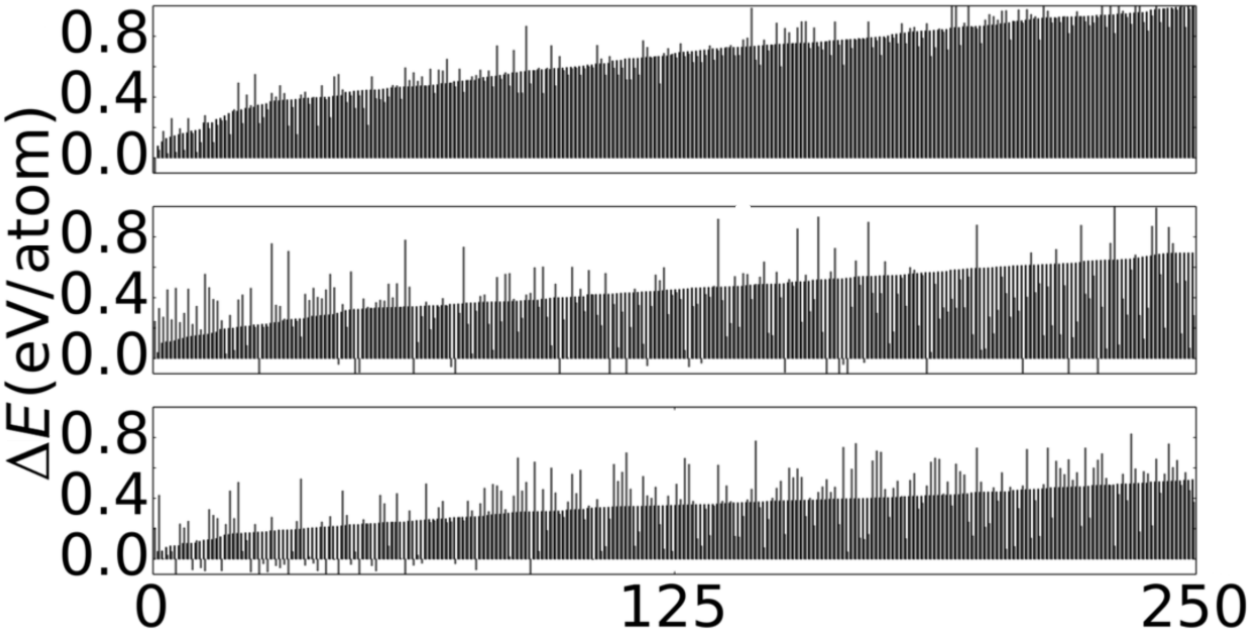



**Figure 4.5:** Lowest 250 ML model predicted formation energies of Elpasolites in ascending order from (III−VI) (TOP) and (I−VIII) (MID and BOTTOM) data sets. Results in TOP and MID panel correspond to ML models trained on 2000 examples, BOTTOM panel results correspond to a ML model trained on 10k crystals. Validating DFT energies are shown aside.

solids other groups report DFT errors on the order of 0.1 eV/atom [210, 211]. The converging performance for training on nearly all crystals of the (III−VI) data set suggests that our crystal representation of Elpasolite structures Fig. 2.1 accounts for the necessary degrees of freedom. While errors decay systematically and linearly on a log-log plot, the learning rate levels off as $N$ approaches the 100%, i.e. 10k. This is due to the employed relaxation convergence threshold of $\pm 10$ meV/atom in the DFT calculations. Any inductive model must fail to go below this level, and only numerically more precise reference numbers would mitigate this trend. In all validation tests dealing with energy predictions for random out-of-sample crystals, the ML model performance meets the expectations set in Fig. 4.5. For example, drawing 100 crystals at random from (III−VI) and (I−VIII) datasets ML models perform as expected when compared to the result from validating DFT calculations (see Fig. 4.4).

(III−VI) and (I−VIII) reaches a MAE of 0.1 eV/atom at roughly 2.5 % and 0.5 % of the total number of crystals respectively, suggesting that the machine "efficiency" increases with number of possible combinations. We note however that two observations of the same structure is not sufficient to see any trends on how much training data is needed.

Having established the performance of the ML model, we have subsequently used the 10 k

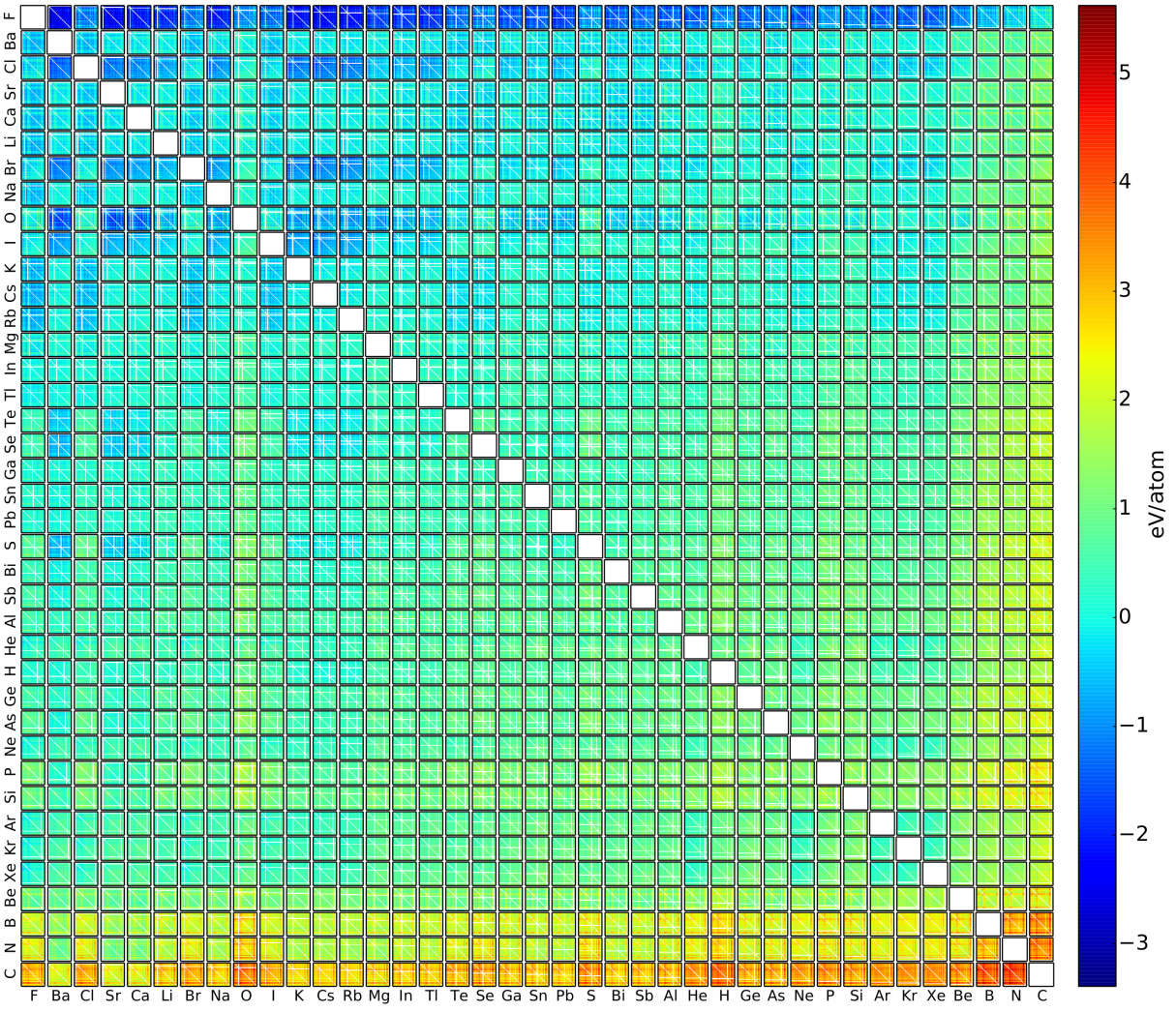



**Figure 4.6:** Formation energies for all 2 M Elpasolites made up of all main-group elements up to Bi predicted by the 10 k ML-model. The outer vertical and horizontal axis correspond to $x_4$ and $x_3$ symmetry position, respectively. Inner vertical and horizontal axis correspond to $x_2$ and $x_1$ symmetry position, respectively. Elemental sequence follows the Elpasolite order of Fig. 4.7. White pixels correspond to subspaces of ternary, binary, or elementary non-Elpasolite crystals.

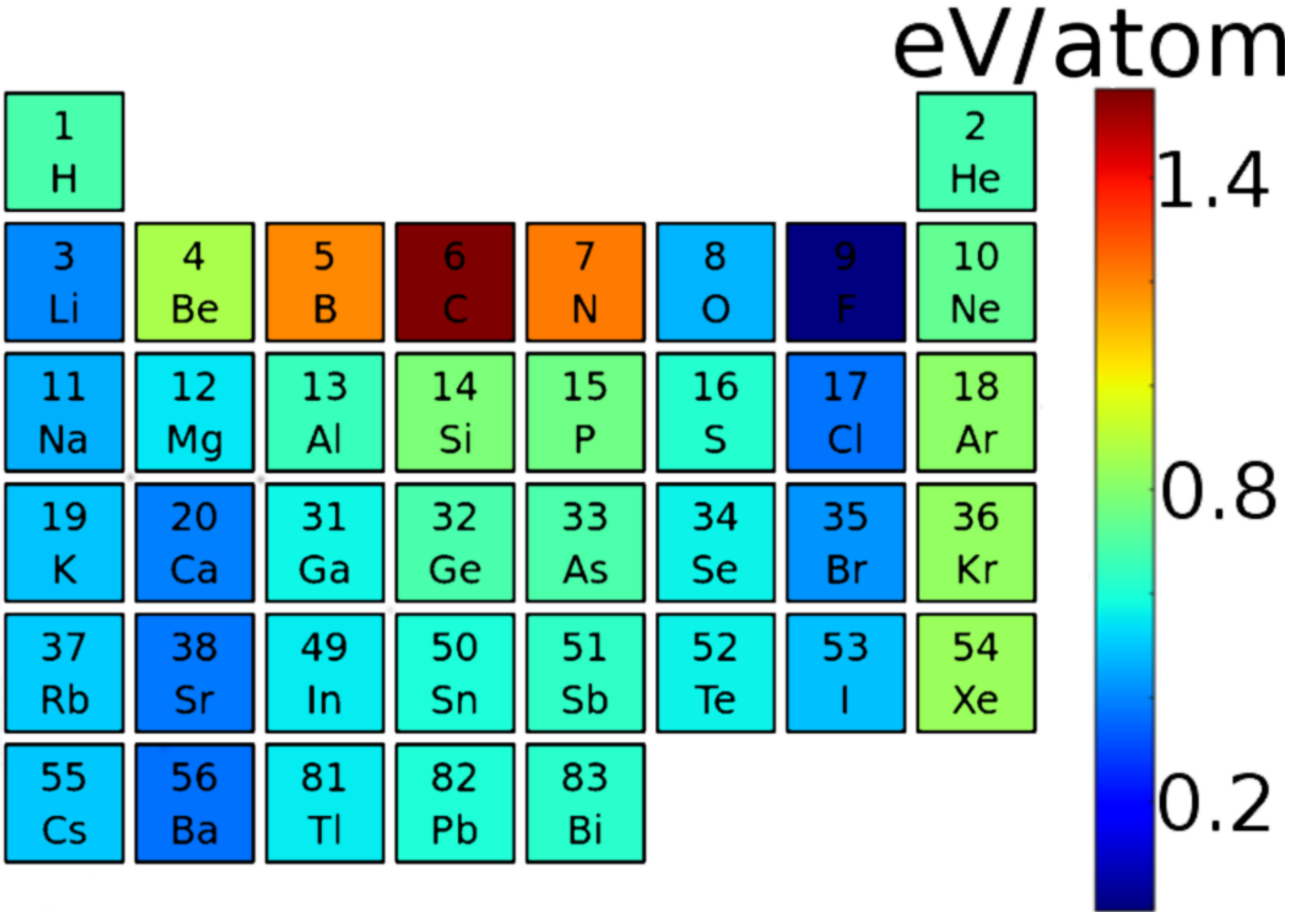



**Figure 4.7:** Estimated mean energy contribution of each element to formation of any Elpasolite crystal. The color code reflects the new elemental Elpasolite order.

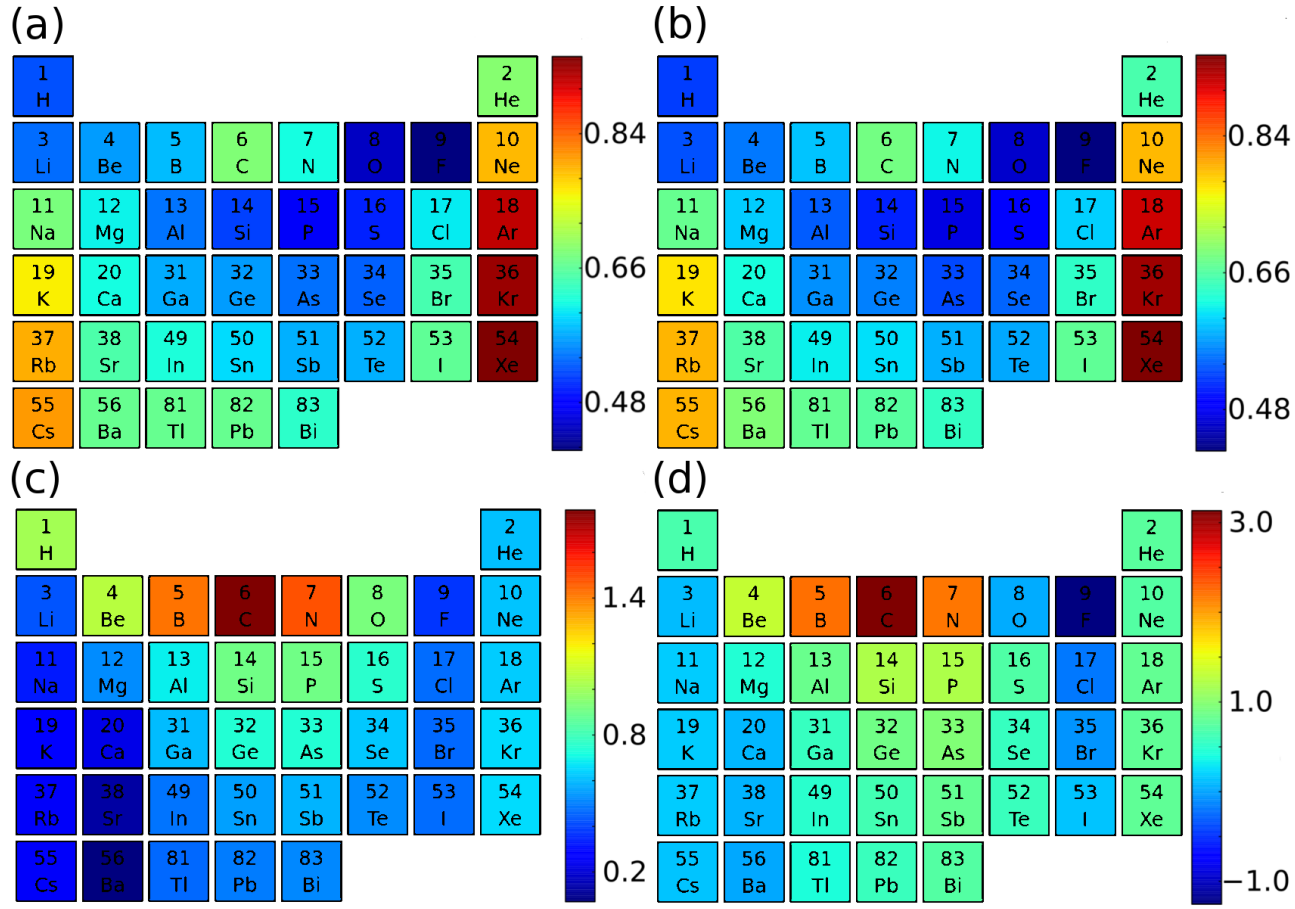



**Figure 4.8:** Site resolved mean contribution to Elpasolite formation energy [eV/atom] for each main-group element. Panels (a), (b), (c), and (d) correspond to respective Elpasolite crystal sites $x_1, x_2, x_3$, and $x_4$ (See Fig. 2.1).

training set model (I−VIII) for investigation of the Elpasolite universe. Estimated formation energies for *all* 2 M Elpasolites are featured in Fig. 4.6. The formation energies are clearly dominated by the chemical identity of position 4, followed by position 3 but according to a different pattern. Chemical identity at position 1 and 2 has the smallest influence and very similar impact, as is illustrated in Fig. 4.8.

Due to the effective degeneracy of positions 1 and 2, all inner matrices in Fig. 4.6 appear largely symmetric. Figure 4.7 shows the average contribution of each element to the formation energies estimated by the 10 k ML model. These average contributions per element are used to order the elements in Fig. 4.6 to yield the smoothest Elpasolite map. Arranging elements by their nuclear charge, or by their Pettifor order [212], results in a much more oscillatory map or stripe-like pattern due to underlying periodicities (see Fig. 4.9).

This Elpasolite error is dominated by the element identity in position 4 (compare Figure 4.7 to Fig. 4.8); its break-down is small as illustrated for pair-wise energy contributions in Fig. 4.10.

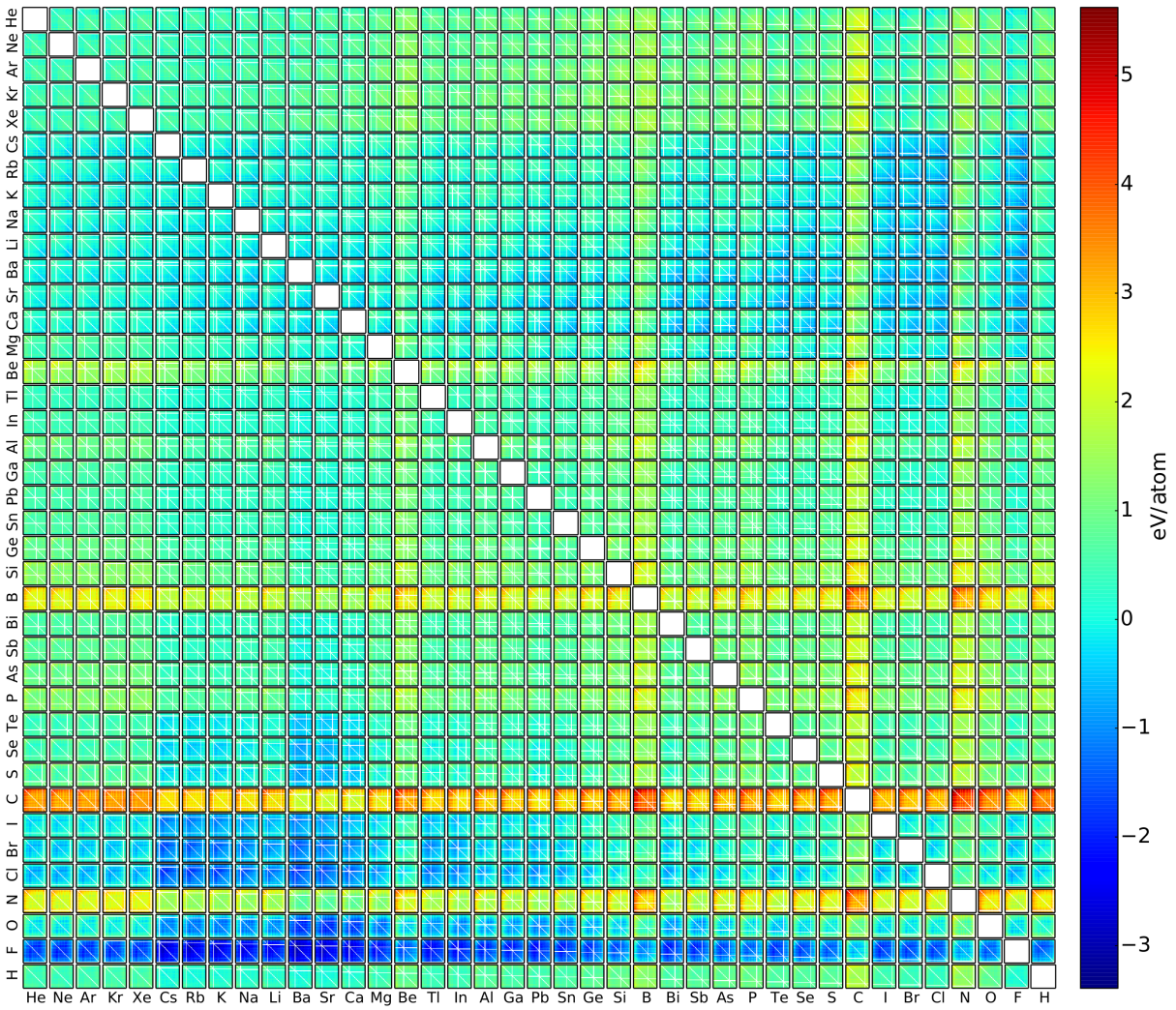



**Figure 4.9:** ML predicted formation energies of all 2 M Elpasolites. The outer vertical axis corresponds to elements in $x_4$ crystal symmetry position, the outer horizontal axis to $x_3$, the inner vertical axis $x_2$ and the inner horizontal axis to the $x_1$ symmetry position (See Fig. 1(a) in the main text). Elemental sequence corresponds to Pettifor order [212]. White lines correspond to ternary, binary, or elemental compositions.

Fluorine and carbon are at the respective ends of the global scale of low and high formation energies. But also alkaline metals, alkaline earth metals, and oxygen contribute to lowering the formation energy. On average, the formation energies of Elpasolites involving halogens, alkaline metals, noble gases increase as the periodic table is descended. The opposite holds for all other elements, except oxygen, boron, carbon and nitrogen, which all have a noticeably higher average formation energy than any other element. A saddle point can also be observed in the midst of the periodic table table as well as two valleys along the halogen and alkaline earth rows. Site-specific resolution indicates that fluorine fits best with the bond coordination of sites 1, 2, and 4, whereas the same does not apply to later halogens, see Fig 4.8. In contrast, as the element on site 3 goes down column II in the periodic table, the formation energy is successively lowered, with Ca, Sr, and Ba contributing more than any halogen atom. On sites 1 and 2, the formation energy generally increases the most for heavy noble gases. On sites 3 and 4, it is carbon, followed by neighboring B and N that increase the formation energy the most. The accuracy of linear single atom energy models based on these scales, however, is not on par with the ML-model, and—maybe more importantly—cannot be improved systematically through increasing training set sizes but rather converges to a finite residual error.

In order to achieve satisfying accuracy of $\pm 0.1$ eV/atom for Elpasolites, a relatively large training set of 10 k is needed. This is likely due to the sparsity of crystals at the opposite ends of the high and low formation energy spectrum; this results in a decreased predictive ML model accuracy for crystals in these regions, which is demonstrated in Fig. 4.12.

Nevertheless, the 10 k ML model readily identifies a larger set of lowest lying Elpasolites for which the actual DFT minima can be obtained through subsequent DFT based screening. This is shown in Fig. 4.11 where the 250 crystals with the lowest ML predicted formation energies are shown in ascending order. Subsequent screening with DFT indicates the $26^{th}$ crystal $CaSrCs_2F_6$ (out of 2M) to be the global formation energy minimum at $-3.44$ eV/atom, closely followed a near-degenerate isomer $SrCaCs_2F_6$. The DFT energies of the next two degenerate pairs $CaSrRb_2F_6$/$SrCaRb_2F_6$ and $CaBaCs_2F_6$/$BaCaCs_2F_6$ correspond to $-3.41$, and $-3.39$ eV/atom, respectively. Overall, the Elpasolites with the most favorable formation energies, $ABC_2D_6$, correspond to A and B being late elements from group (II), and C and D being a late element from group (I) and fluoride, respectively. Populating the four sites with elements from groups (II),(II),(I), and (VIII), respectively, differs from the experimentally established stoichiometry $AlNaK_2F_6$. In fact, the lowest DFT energy crystal with a group-(III) element is $CsAlRb_2F_6$ (in $69^{th}$ position) with $-3.09$ eV/atom (ML energy: $-2.96$ eV/atom, see Table 4.2). We

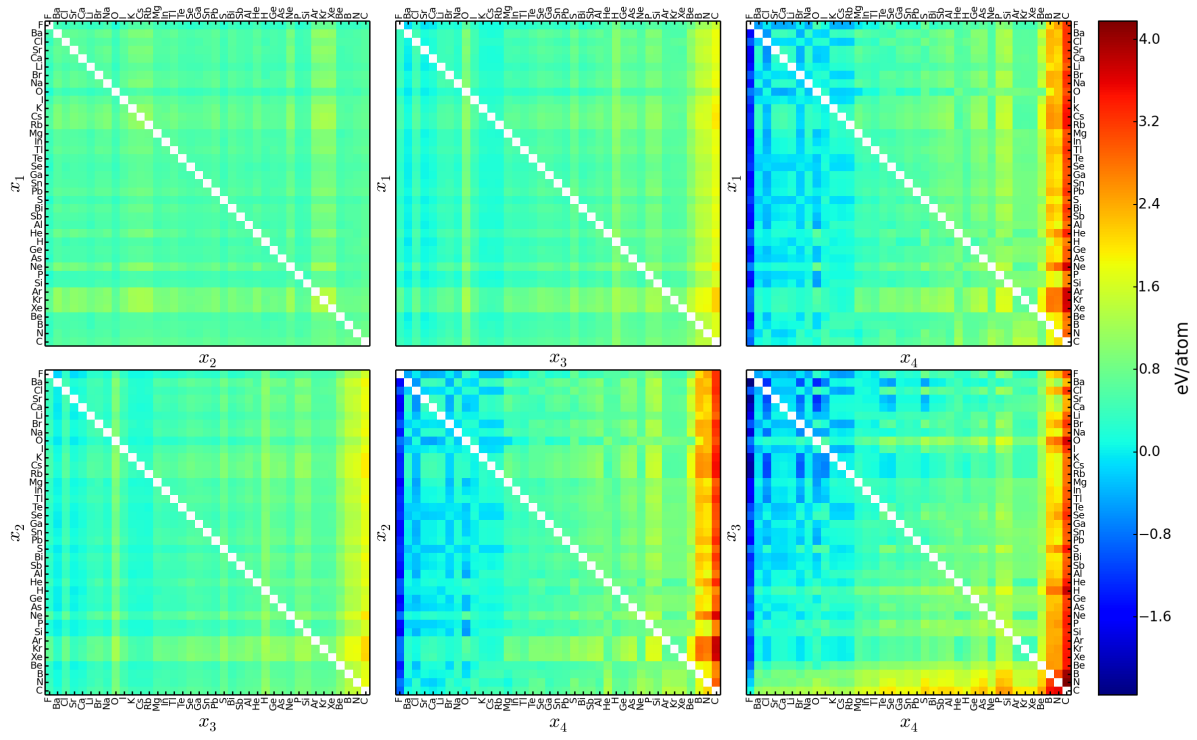



**Figure 4.10:** Average two-body formation energy for any combination of placing two elements at two different sites in the four-tuple $x$ (See Fig. 2.1). Elemental sequence follows the Elpasolite energy order (See Fig. 4.7).

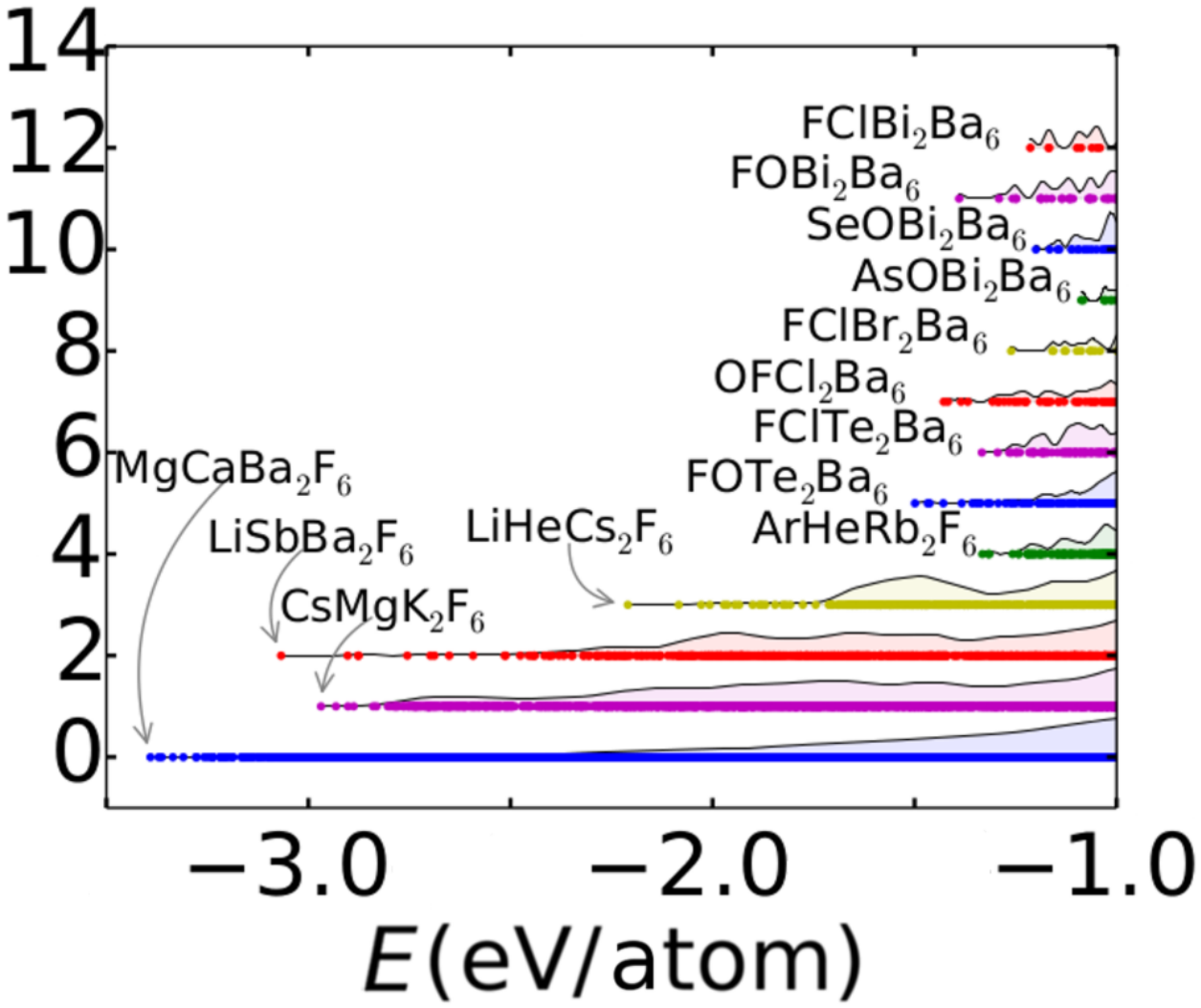



**Figure 4.11:** Distributions of LPTOS in energies. Formulas indicate the lowest lying crystals.

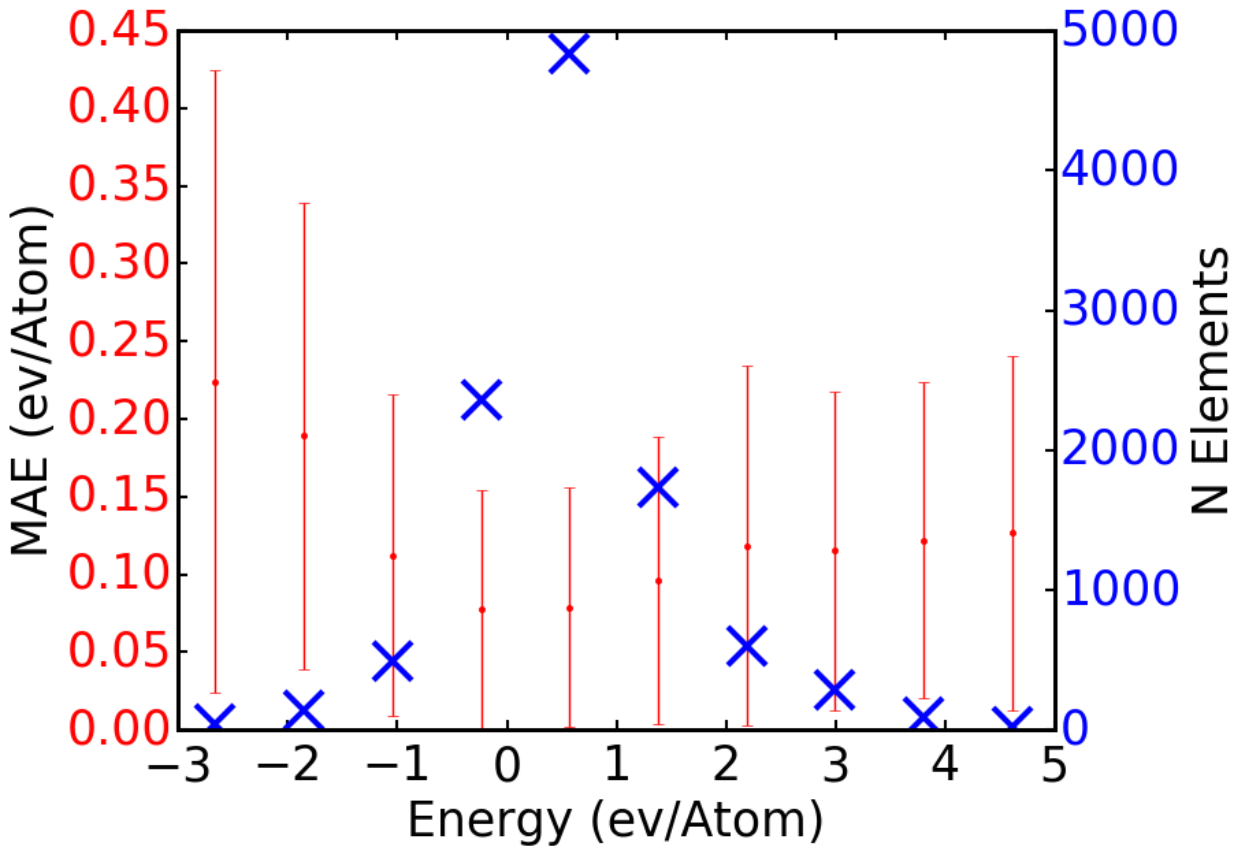



**Figure 4.12:** Number of crystals per energy bin correlates inversely with out-of-sample prediction error. (Left vertical axis) Out-of-sample MAE distribution in DFT formation energy of 10'590 crystals (training and test sets). Standard deviations have been obtained for 100 different ML models trained on randomly chosen 10k reference crystals. (Right vertical axis) Number of crystal structures per formation energy bin.

have also used our predictions to analyse atomic oxidation states in Elpasolites. In particular, we have found that roughly 6 % of the crystals with formation energies below $-1$ eV/atom exhibit unusual atomic charges: They are low in energy despite the fact that no combination of conventional atomic charges would result in a neutral system. In order to identify these crystals, we have used the absolute value of the lowest possible total oxidation state (LPTOS) that could possibly be realized using the list of typical atomic oxidation states on display in Table 4.1.

The lowest lying crystals have a LPTOS of 0 ($-3$ to $-3.44$ eV/atom formation energies). However, already at $-3$ eV/atom crystals with LPTOS of 2 or 1 start to occur. At formation energies of $\sim -1.25$ eV/atom and higher, the number of crystals with non-zero LPTOS increases rapidly, with LPTOS as high as 12. Corresponding crystal frequency distributions are shown in

**Table 4.1:** Conventional oxidation states for all elements considered in this work (values taken from `wikipedia.org`). Unconventional oxidation states found in this study are highlighted in red.

| Element | -5 | -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | | | | | ✓ | | ✓ | | | | | | |
| He | | | | | | ✓ | | | | | | | |
| Li | | | | | | | ✓ | | | | | | |
| Be | | | | | | | ✓ | ✓ | | | | | |
| B | ✓ | | | | | | ✓ | ✓ | ✓ | | | | |
| C | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | |
| N | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| O | | | | ✓ | ✓ | | ✓ | ✓ | | | | | |
| F | | | | | ✓ | | | | | | | | |
| Ne | | | | | | ✓ | | | | | | | |
| Na | | | | | ✓ | | ✓ | | | | | | |
| Mg | | | | | | | ✓ | ✓ | | | | | |
| Al | | | | | | | ✓ | ✓ | ✓ | | | | |
| Si | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | | | |
| P | | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| S | | | | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Cl | | | | | ✓ | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Ar | | | | | | ✓ | | | | | | | |
| K | | | | | ✓ | | ✓ | | | | | | |
| Ca | | | | | ✓ | | ✓ | ✓ | | | | | |

| Element | -5 | -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ga | ✓ | ✓ | | ✓ | | | ✓ | ✓ | ✓ | | | | | |
| Ge | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | |
| As | | | ✓ | | | | ✓ | ✓ | ✓ | | ✓ | | | |
| Se | | | | ✓ | | | ✓ | ✓ | | ✓ | | ✓ | | |
| Br | | | | | ✓ | | ✓ | | ✓ | ✓ | ✓ | | ✓ | |
| Kr | | | | | | ✓ | ✓ | ✓ | | | | | | |
| Rb | | | | | ✓ | | ✓ | | | | | | | |
| Sr | | | | | | | ✓ | ✓ | | | | | | |
| In | ✓ | | | | | | ✓ | ✓ | ✓ | | | | | |
| Sn | | ✓ | | | | | ✓ | ✓ | ✓ | ✓ | | | | |
| Sb | | | ✓ | | | ✓ | ✓ | | ✓ | | ✓ | | | |
| Te | | | | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | |
| I | | | | | ✓ | | ✓ | | | ✓ | | ✓ | ✓ | |
| Xe | | | | | | | ✓ | ✓ | | ✓ | | ✓ | | ✓ |
| Cs | | | | | ✓ | | ✓ | | | | | | | |
| Ba | | | | | | | | ✓ | | | | | | |
| Ti | ✓ | | | | | | ✓ | ✓ | ✓ | | | | | |
| Pb | | | | | | | ✓ | ✓ | ✓ | ✓ | | | | |
| Bi | | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | | | |

Fig. 4.11, along with formulas for the mutually lowest lying crystals. Interestingly, the number of crystals with zero LPTOS increases monotonically with formation energy, while for nonzero LPTOS crystals the distribution is oscillatory.

To demonstrate the usefulness of our ML model we have applied it to identify thermodynamically stable Elpasolites. To this end, we first selected all those 274,213 Elpasolites with negative ML formation energies, and without rare gas elements. Since stability depends on the energy difference to any possible polymorph or competing segregated phases [213, 214], we have queried available DFT formation energies stored in the Materials Project (MP) [208]. Some Elpasolites, such as the archetypical $AlNaK_2F_6$, are already stored in the MP.

Subsequently, we have used the software package `pymatgen` [215, 216] to create a phase diagram for each of those systems, comparing ML energies to all relevant competing pure or mixed stable phases in the materials project (MP) database. There are two technicalities in this comparison: Our DFT calculations and those of the MP are not done with the exact same set of input parameters. The results reported below have been obtained by adjusting the ML formation energies before using them in the MP phase diagrams: We add the appropriate fraction of the DFT energy difference of pure elemental phases calculated by us and by MP. We also turn off all MP corrections except for the 'MP gas corrections' that are applied to pure elemental phases. However, these remaing MP gas corrections should not affect the results much, since they apply equally to ours and MPs results (and, in addition, they cannot easily be disregarded

**Table 4.2:** 250 Elpasolite crystals predicted by the 10 k ML-model to be the lowest in formation energy, as shown in lower panel of Fig. 1 (e). ML formation energies [eV/atom] are shown together with their corresponding DFT energies [eV/atom], together with the index that sorts the DFT energies.

| Formula | # | ML | DFT | Formula | # | ML | DFT | Formula | # | ML | DFT | Formula | # | ML | DFT | Formula | # | ML | DFT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $MgCaBa_2F_6$ | 70 | -3.27 | -3.07 | $AlSrRb_2F_6$ | 94 | -3.04 | -3.01 | $CaGeBa_2F_6$ | 201 | -2.95 | -2.72 | $CaSnBa_2F_6$ | 189 | -2.89 | -2.76 | $BaBeCs_2F_6$ | 82 | -2.83 | -3.04 |
| $AlCaBa_2F_6$ | 147 | -3.22 | -2.85 | $CaMgSr_2F_6$ | 81 | -3.04 | -3.04 | $NaAlBa_2F_6$ | 152 | -2.95 | -2.84 | $CaBaNa_2F_6$ | 66 | -2.89 | -3.11 | $TlSrBa_2F_6$ | 169 | -2.83 | -2.8 |
| $BaMgCs_2F_6$ | 30 | -3.22 | -3.27 | $SrMgCs_2F_6$ | 18 | -3.04 | -3.32 | $AlBeBa_2F_6$ | 212 | -2.95 | -2.69 | $CaAsBa_2F_6$ | 217 | -2.89 | -2.68 | $CaInBa_2F_6$ | 134 | -2.83 | -2.89 |
| $BaMgRb_2F_6$ | 35 | -3.2 | -3.25 | $CaMgRb_2F_6$ | 10 | -3.04 | -3.35 | $AlCaNa_2F_6$ | 93 | -2.95 | -3.01 | $BaSiCs_2F_6$ | 196 | -2.89 | -2.74 | $AlSrCa_2F_6$ | 206 | -2.83 | -2.71 |
| $BaMgK_2F_6$ | 38 | -3.19 | -3.23 | $CaAlBa_2F_6$ | 148 | -3.04 | -2.85 | $SrMgCa_2F_6$ | 122 | -2.94 | -2.92 | $CaSbBa_2F_6$ | 214 | -2.89 | -2.68 | $CaLiSr_2F_6$ | 44 | -2.83 | -3.21 |
| $BaCaCs_2F_6$ | 5 | -3.19 | -3.39 | $GaCaBa_2F_6$ | 124 | -3.03 | -2.91 | $CaLiBa_2F_6$ | 29 | -2.94 | -3.28 | $BaGaK_2F_6$ | 197 | -2.89 | -2.74 | $LiAlCa_2F_6$ | 170 | -2.83 | -2.8 |
| $MgCaSr_2F_6$ | 80 | -3.19 | -3.04 | $NaMgBa_2F_6$ | 50 | -3.03 | -3.19 | $AlLiSr_2F_6$ | 139 | -2.93 | -2.88 | $CsCaBa_2F_6$ | 106 | -2.89 | -2.99 | $LiSrBa_2F_6$ | 47 | -2.82 | -3.19 |
| $CaMgBa_2F_6$ | 71 | -3.17 | -3.07 | $BaCaSr_2F_6$ | 86 | -3.03 | -3.03 | $MgCaNa_2F_6$ | 42 | -2.93 | -3.22 | $AlKBa_2F_6$ | 141 | -2.89 | -2.87 | $InLiBa_2F_6$ | 118 | -2.82 | -2.94 |
| $LiBeBa_2F_6$ | 90 | -3.17 | -3.03 | $MgBeBa_2F_6$ | 151 | -3.03 | -2.84 | $SrLiBa_2F_6$ | 48 | -2.93 | -3.19 | $SrBaNa_2F_6$ | 75 | -2.89 | -3.06 | $CaGaSr_2F_6$ | 175 | -2.82 | -2.79 |
| $BaCaRb_2F_6$ | 15 | -3.17 | -3.34 | $MgBaCs_2F_6$ | 31 | -3.03 | -3.27 | $NaBeBa_2F_6$ | 96 | -2.93 | -3.01 | $MgBaCa_2F_6$ | 156 | -2.88 | -2.83 | $SrTeBa_2F_6$ | 241 | -2.82 | -2.55 |
| $SrCaBa_2F_6$ | 53 | -3.17 | -3.15 | $CaSrRb_2F_6$ | 2 | -3.03 | -3.41 | $BLiBa_2F_6$ | 223 | -2.93 | -2.65 | $AlLiCa_2F_6$ | 173 | -2.88 | -2.8 | $SrNaBa_2F_6$ | 58 | -2.82 | -3.14 |
| $MgCaCs_2F_6$ | 17 | -3.16 | -3.33 | $AlLiBa_2F_6$ | 113 | -3.02 | -2.96 | $BaGaRb_2F_6$ | 188 | -2.93 | -2.76 | $TlSrCs_2F_6$ | 193 | -2.88 | -2.75 | $NaGaBa_2F_6$ | 135 | -2.81 | -2.89 |
| $MgCaRb_2F_6$ | 11 | -3.15 | -3.35 | $SrMgRb_2F_6$ | 21 | -3.02 | -3.31 | $MgBeCa_2F_6$ | 209 | -2.93 | -2.7 | $BaMgCa_2F_6$ | 155 | -2.88 | -2.83 | $CaKBa_2F_6$ | 62 | -2.81 | -3.12 |
| $AlCaCs_2F_6$ | 114 | -3.15 | -2.95 | $MgBaRb_2F_6$ | 34 | -3.02 | -3.25 | $CsMgK_2F_6$ | 238 | -2.93 | -2.58 | $MgAlBa_2F_6$ | 172 | -2.88 | -2.8 | $LiGeBa_2F_6$ | 180 | -2.81 | -2.78 |
| $BaAlCs_2F_6$ | 108 | -3.15 | -2.99 | $SrBaRb_2F_6$ | 26 | -3.02 | -3.3 | $GaSrBa_2F_6$ | 165 | -2.93 | -2.81 | $MgBaNa_2F_6$ | 76 | -2.88 | -3.05 | $SiSrBa_2F_6$ | 243 | -2.81 | -2.54 |
| $BaAlK_2F_6$ | 98 | -3.14 | -3.01 | $BeMgBa_2F_6$ | 183 | -3.02 | -2.78 | $LiMgSr_2F_6$ | 49 | -2.93 | -3.19 | $SiCaSr_2F_6$ | 245 | -2.88 | -2.54 | $LiGaBa_2F_6$ | 100 | -2.8 | -3.0 |
| $BaSrCs_2F_6$ | 12 | -3.13 | -3.35 | $SrAlK_2F_6$ | 84 | -3.01 | -3.03 | $KMgBa_2F_6$ | 67 | -2.93 | -3.1 | $CaNaBa_2F_6$ | 36 | -2.88 | -3.23 | $BeAsBa_2F_6$ | 227 | -2.8 | -2.63 |
| $MgSrBa_2F_6$ | 79 | -3.11 | -3.05 | $LiBeSr_2F_6$ | 91 | -3.01 | -3.02 | $AlMgSr_2F_6$ | 198 | -2.93 | -2.73 | $SrSnBa_2F_6$ | 213 | -2.88 | -2.68 | $CaTeBa_2F_6$ | 216 | -2.8 | -2.68 |
| $BeCaBa_2F_6$ | 158 | -3.11 | -2.83 | $BaMgSr_2F_6$ | 119 | -3.01 | -2.94 | $InSrCs_2F_6$ | 146 | -2.93 | -2.86 | $BCaBa_2F_6$ | 247 | -2.88 | -2.51 | $LiSiBa_2F_6$ | 200 | -2.8 | -2.72 |
| $AlBaK_2F_6$ | 97 | -3.11 | -3.01 | $AlMgK_2F_6$ | 130 | -3.01 | -2.89 | $KAlBa_2F_6$ | 140 | -2.92 | -2.87 | $NaSrBa_2F_6$ | 59 | -2.88 | -3.14 | $LiSbBa_2F_6$ | 231 | -2.8 | -2.62 |
| $AlSrBa_2F_6$ | 184 | -3.1 | -2.77 | $AlSrK_2F_6$ | 83 | -3.01 | -3.03 | $InSrBa_2F_6$ | 171 | -2.92 | -2.8 | $SrMgNa_2F_6$ | 55 | -2.87 | -3.15 | $AlGaBa_2F_6$ | 192 | -2.8 | -2.75 |
| $LiMgBa_2F_6$ | 41 | -3.1 | -3.22 | $CaAlK_2F_6$ | 88 | -3.0 | -3.03 | $BaMgNa_2F_6$ | 77 | -2.92 | -3.05 | $BaCaNa_2F_6$ | 65 | -2.87 | -3.11 | $BaSiRb_2F_6$ | 207 | -2.8 | -2.71 |
| $CaSrBa_2F_6$ | 54 | -3.1 | -3.15 | $MgSrK_2F_6$ | 24 | -3.0 | -3.3 | $LiCaSr_2F_6$ | 45 | -2.92 | -3.21 | $SrGeBa_2F_6$ | 228 | -2.87 | -2.63 | $RbCaBa_2F_6$ | 73 | -2.8 | -3.06 |
| $CaMgK_2F_6$ | 8 | -3.1 | -3.36 | $AlBaRb_2F_6$ | 101 | -3.0 | -3.0 | $CaGaCs_2F_6$ | 145 | -2.92 | -2.86 | $SrSbBa_2F_6$ | 240 | -2.87 | -2.56 | $SrSiBa_2F_6$ | 244 | -2.8 | -2.54 |
| $SrMgBa_2F_6$ | 78 | -3.1 | -3.05 | $BeLiSr_2F_6$ | 92 | -3.0 | -3.02 | $CaMgNa_2F_6$ | 43 | -2.92 | -3.22 | $LiPBa_2F_6$ | 239 | -2.87 | -2.57 | $GaLiBa_2F_6$ | 99 | -2.79 | -3.0 |
| $CaSrCs_2F_6$ | 0 | -3.1 | -3.44 | $InCaBa_2F_6$ | 133 | -3.0 | -2.89 | $TlCaCs_2F_6$ | 182 | -2.92 | -2.78 | $BeSrBa_2F_6$ | 163 | -2.87 | -2.82 | $BaSiK_2F_6$ | 219 | -2.79 | -2.67 |
| $SrMgK_2F_6$ | 25 | -3.1 | -3.3 | $LiAlBa_2F_6$ | 112 | -3.0 | -2.96 | $TlMgBa_2F_6$ | 138 | -2.92 | -2.88 | $BeLiCa_2F_6$ | 104 | -2.87 | -3.0 | $CaGeSr_2F_6$ | 230 | -2.79 | -2.62 |
| $MgCaK_2F_6$ | 9 | -3.09 | -3.36 | $CaBaRb_2F_6$ | 14 | -3.0 | -3.34 | $CsMgRb_2F_6$ | 234 | -2.92 | -2.61 | $CaBeSr_2F_6$ | 167 | -2.87 | -2.81 | $BLiSr_2F_6$ | 237 | -2.79 | -2.58 |
| $BaAlRb_2F_6$ | 103 | -3.09 | -3.0 | $CaGaBa_2F_6$ | 125 | -2.99 | -2.91 | $LiBBa_2F_6$ | 222 | -2.92 | -2.65 | $SnCaBa_2F_6$ | 190 | -2.87 | -2.76 | $MgSnBa_2F_6$ | 195 | -2.79 | -2.74 |
| $MgSrCs_2F_6$ | 19 | -3.09 | -3.32 | $BeCaSr_2F_6$ | 168 | -2.99 | -2.81 | $GaMgBa_2F_6$ | 131 | -2.92 | -2.89 | $AlNaBa_2F_6$ | 143 | -2.87 | -2.86 | $MgGaSr_2F_6$ | 179 | -2.79 | -2.79 |
| $CaMgCs_2F_6$ | 16 | -3.08 | -3.33 | $CsMgBa_2F_6$ | 111 | -2.99 | -2.96 | $NaCaSr_2F_6$ | 56 | -2.91 | -3.14 | $AlRbCs_2F_6$ | 57 | -2.86 | -3.14 | $MgNaBa_2F_6$ | 51 | -2.79 | -3.19 |
| $BaCaK_2F_6$ | 22 | -3.08 | -3.31 | $MgBeSr_2F_6$ | 181 | -2.98 | -2.78 | $MgLiSr_2F_6$ | 52 | -2.91 | -3.19 | $AlBaSr_2F_6$ | 191 | -2.86 | -2.76 | $AlRbBa_2F_6$ | 149 | -2.78 | -2.85 |
| $SrCaCs_2F_6$ | 1 | -3.08 | -3.44 | $BaGaCs_2F_6$ | 177 | -2.98 | -2.79 | $AlBaNa_2F_6$ | 117 | -2.91 | -2.95 | $CaAlSr_2F_6$ | 205 | -2.86 | -2.71 | $GeCaBa_2F_6$ | 202 | -2.78 | -2.72 |
| $MgBaK_2F_6$ | 39 | -3.08 | -3.23 | $CaBeBa_2F_6$ | 160 | -2.98 | -2.83 | $LiMgCa_2F_6$ | 64 | -2.91 | -3.12 | $BaAlNa_2F_6$ | 116 | -2.86 | -2.95 | $GaNaBa_2F_6$ | 136 | -2.78 | -2.89 |
| $BeLiBa_2F_6$ | 89 | -3.08 | -3.03 | $CaAlCs_2F_6$ | 115 | -2.97 | -2.95 | $HBeBa_2F_6$ | 121 | -2.91 | -2.93 | $MgAlK_2F_6$ | 126 | -2.86 | -2.91 | $BBeBa_2F_6$ | 249 | -2.78 | -2.45 |
| $AlCaSr_2F_6$ | 194 | -3.08 | -2.75 | $BaSrK_2F_6$ | 33 | -2.97 | -3.26 | $TlMgCs_2F_6$ | 221 | -2.91 | -2.66 | $GaCaSr_2F_6$ | 174 | -2.86 | -2.79 | $SbCaBa_2F_6$ | 215 | -2.78 | -2.68 |
| $MgSrRb_2F_6$ | 20 | -3.07 | -3.31 | $GaCaCs_2F_6$ | 144 | -2.97 | -2.86 | $MgGaBa_2F_6$ | 132 | -2.91 | -2.89 | $CaSiBa_2F_6$ | 225 | -2.86 | -2.64 | $MgKBa_2F_6$ | 68 | -2.78 | -3.1 |
| $SrBaK_2F_6$ | 32 | -3.07 | -3.26 | $AlMgCa_2F_6$ | 236 | -2.97 | -2.61 | $GaBaCs_2F_6$ | 176 | -2.91 | -2.79 | $AlBeSr_2F_6$ | 235 | -2.86 | -2.61 | $MgSbBa_2F_6$ | 211 | -2.77 | -2.7 |
| $BaSrRb_2F_6$ | 27 | -3.07 | -3.3 | $CsAlBa_2F_6$ | 157 | -2.96 | -2.83 | $AlMgNa_2F_6$ | 129 | -2.91 | -2.9 | $LiAsBa_2F_6$ | 232 | -2.85 | -2.62 | $MgAsBa_2F_6$ | 204 | -2.77 | -2.71 |
| $SrBaCs_2F_6$ | 13 | -3.07 | -3.35 | $SrAlBa_2F_6$ | 185 | -2.96 | -2.77 | $KCaBa_2F_6$ | 61 | -2.9 | -3.12 | $PbMgK_2F_6$ | 150 | -2.85 | -2.85 | $BaSrNa_2F_6$ | 74 | -2.77 | -3.06 |
| $AlCaK_2F_6$ | 87 | -3.06 | -3.03 | $CaSrK_2F_6$ | 6 | -2.96 | -3.38 | $NaMgSr_2F_6$ | 60 | -2.9 | -3.13 | $GaMgK_2F_6$ | 187 | -2.85 | -2.77 | $BeGeBa_2F_6$ | 233 | -2.77 | -2.62 |
| $SrCaRb_2F_6$ | 3 | -3.06 | -3.41 | $SiCaBa_2F_6$ | 226 | -2.96 | -2.64 | $CaAlRb_2F_6$ | 107 | -2.9 | -2.99 | $MgSrCa_2F_6$ | 123 | -2.85 | -2.92 | $LiAlSr_2F_6$ | 153 | -2.76 | -2.84 |
| $AlCaRb_2F_6$ | 105 | -3.06 | -2.99 | $MgLiBa_2F_6$ | 40 | -2.96 | -3.22 | $InCaCs_2F_6$ | 137 | -2.9 | -2.88 | $TlMgK_2F_6$ | 224 | -2.85 | -2.65 | $MgAlSr_2F_6$ | 203 | -2.76 | -2.71 |
| $LiCaBa_2F_6$ | 28 | -3.06 | -3.28 | $SrBeBa_2F_6$ | 162 | -2.96 | -2.82 | $SrGaBa_2F_6$ | 164 | -2.9 | -2.81 | $BeMgSr_2F_6$ | 210 | -2.85 | -2.7 | $CaSbSr_2F_6$ | 246 | -2.76 | -2.52 |
| $SrCaK_2F_6$ | 7 | -3.06 | -3.38 | $CsAlRb_2F_6$ | 69 | -2.96 | -3.09 | $AlBeCa_2F_6$ | 248 | -2.9 | -2.5 | $MgGeBa_2F_6$ | 199 | -2.84 | -2.72 | $CaSnSr_2F_6$ | 220 | -2.76 | -2.67 |
| $AlMgBa_2F_6$ | 159 | -3.06 | -2.83 | $SiBaK_2F_6$ | 218 | -2.96 | -2.68 | $MgBaSr_2F_6$ | 120 | -2.9 | -2.94 | $CaBaSr_2F_6$ | 85 | -2.84 | -3.03 | $BeAlBa_2F_6$ | 229 | -2.76 | -2.62 |
| $AlBaCs_2F_6$ | 109 | -3.05 | -2.99 | $GaSrCs_2F_6$ | 154 | -2.96 | -2.84 | $SrCaNa_2F_6$ | 46 | -2.9 | -3.2 | $BeCaRb_2F_6$ | 63 | -2.84 | -3.12 | $GaMgSr_2F_6$ | 178 | -2.76 | -2.79 |
| $CaBaK_2F_6$ | 23 | -3.05 | -3.31 | $SrAlCs_2F_6$ | 110 | -2.96 | -2.98 | $SrAlRb_2F_6$ | 95 | -2.89 | -3.01 | $SrAsBa_2F_6$ | 242 | -2.83 | -2.54 | $GaKBa_2F_6$ | 208 | -2.75 | -2.7 |
| $NaCaBa_2F_6$ | 37 | -3.05 | -3.23 | $LiBeCa_2F_6$ | 102 | -2.96 | -3.0 | $InMgBa_2F_6$ | 142 | -2.89 | -2.87 | $InCaSr_2F_6$ | 186 | -2.83 | -2.77 | $AlKSr_2F_6$ | 161 | -2.75 | -2.82 |
| $CaBaCs_2F_6$ | 4 | -3.05 | -3.39 | $TlCaBa_2F_6$ | 128 | -2.95 | -2.9 | $BaInCs_2F_6$ | 166 | -2.89 | -2.81 | $BeCaCs_2F_6$ | 72 | -2.83 | -3.06 | $LiHBa_2F_6$ | 127 | -2.75 | -2.9 |

when querying MP for phase diagram data.) We have also tried other options (i.e., comparing absolute energies and leaving in the MP oxide corrections), with some additional findings on the differences also reported below.

Using our ML predictions in combination with pre-existing data from the MP, we have identified 2133 configurations (out of ∼0.3M with negative formation energies) which are predicted to reside below the convex hull of stability. Such a major reduction suggests the MP database of competing phases to be important for our analysis to be meaningful. The phase diagrams were found to have on average ∼12 competing phases in addition to the elemental phases. Using DFT we have validated all the 2133 crystals, and 128 are confirmed to be below the convex hull. Since the $ABC_2D_6$ and $BAC_2D_6$ Elpasolite systems are energetically very similar, some of the configurations found correspond to such polymorphs. Discounting polymorphs, we predict 90 Elpasolite systems below the convex hull (shown in Table 4.3).

Other choices of comparing our ML energies to MP entries, and, adding the MP oxide correction, add 6 more systems predicted below the hull, 3 of them confirmed by DFT ($BrICs_2Cl_6$, $IBrCs_2Cl_6$ and $SbNaCs_2Cl_6$. These 3 additions are just slightly below the hull with $< 5$ meV/atom. Note that it is likely that some of these systems can be even further relaxed below the hull if one allows for breaking the Elpasolite symmetry. We have submitted these systems to MP for confirmation (using exactly equivalent DFT settings) where the vast majority of our Elpasolites has been confirmed to be stable, and has been included. At the time of writing, $InCsRb_2F_6$ ,$GaNaCs_2H_6$, $TlGaCs_2H_6$ and $GaAlCs_2H_6$ were found not to lie on the convex hull.

The large discrepancy between configurations predicted by ML and confirmed stable by DFT is expected since the selection of ML energies below the convex hull systematically promotes systems that have negative energy errors. To clarify, our selection of 2133 configurations with ML energy below the hull does not just include the 128 that we later confirm by DFT, but *any* configuration where the error in the predicted ML energy is negative and large enough to bring the error below the hull. For example, any large negative outliers in the complete set of considered Elpasolites are included in this set, as well as, smaller error outliers more closely above the hull. Still, it turns out this is not a problem of major practical significance since the ML accuracy apparently is good enough for only roughly 1 out of 140 entries (2000 out of 275k) to be incorrectly brought down below the hull. Hence, this error rate is still small enough to allow exhaustive validation of all found entries to dismiss those incorrectly included.

We also note that this does not amount to proof that the 90 crystals are stable: The MP database is not exhaustive. This implies that other new competing phases and materials, with

even stronger stabilization, might still be discovered in the future. Also, the intrinsic error of the employed DFT method within the MP might still alter the outcome with respect to experiment. As such, the 90 new Elpasolite DFT energies represent new upper bounds on the convex hull at the corresponding compositions. They have been submitted to the MP database, and most of them have been made available for further studies (See Table 4.3 for a list of the 90 structures).

**Table 4.3:** List of 90 Elpasolites discovered in this study, in the order of calculated increasing stability, with ML predicted formation energies below the presently known convex hull in the MP database that subsequently have also been confirmed by DFT calculations to reside on the convex hull. $\Delta E$ indicates the decrease in DFT energy with respect to the formation energy of the lowest previously known energy of the 'competing mixed phase' in the MP database . When available, the MP reference Id (MP-ID) is listed, together with the MP calculated band-gap($\Delta\epsilon$), and a type tag: conductor (C) if 0.1 eV $\geq \Delta\epsilon$, semiconductor (S) if $\Delta\epsilon > 3$ eV and insulator (I) if 3.0 eV $\geq \Delta\epsilon > 0.1$ eV based on the MP band-gap: conductor (0.1 eV $\geq \Delta\epsilon$).

| # | Elpasolite $x_1$ $x_2$ $x_3$ $x_4$ | $\Delta E$ (eV/atom) | Type | $\Delta\epsilon$ (eV) | MP ID | Competing mixed phase in MP |
|---|---|---|---|---|---|---|
| 1 | Li Tl Rb Cl | −0.0003 | S | 1.56 | mp-989579 | $0.1667\,Rb_{12}Tl_4Cl_{24} + 0.0833\,Tl_4Cl_8 + 0.0833\,Cl_4 + Li_1Cl_1$ |
| 2 | Sb Na Rb F | −0.0011 | I | 4.43 | mp-989541 | $0.125\,Na_8Sb_4F_{20} + 0.0625\,Rb_4Sb_8F_{28} + 1.75\,Rb_1F_1$ |
| 3 | Ga Cs Rb F | −0.0012 | I | 5.54 | mp-989629 | $0.0556\,Cs_{18}Ga_{12}F_{54} + 0.1667\,Ga_2F_6 + 2\,Rb_1F_1$ |
| 4 | N F In Sr | −0.0015 | C | 0.00 | mp-989402 | $0.4853\,Sr_2N_1 + 0.5\,Sr_1F_2 + 0.2574\,Sr_8In_4N_2 + 0.0882\,Sr_{28}In_{11}$ |
| 5 | N P Bi Mg | −0.0021 | S | 0.23 | mp-989522 | $0.0625\,Mg_{24}P_{16} + Mg_3Bi_2 + 0.0625\,Mg_{24}N_{16}$ |
| 6 | In Cs Rb F | −0.0027 | I | 3.73 | mp-989595 | $0.0833\,Rb_8In_{12}F_{44} + 1.3333\,Rb_1F_1 + Cs_1F_1$ |
| 7 | As Br Cs Cl | −0.0027 | C | 0.07 | mp-989511 | $0.25\,As_4Cl_{12} + 0.5\,Br_2Cl_2 + 2\,Cs_1Cl_1$ |
| 8 | Tl Na Rb Cl | −0.0030 | S | 1.81 | mp-989563 | $0.1667\,Rb_{12}Tl_4Cl_{24} + 0.0833\,Tl_4Cl_8 + 0.0833\,Cl_4 + Na_1Cl_1$ |
| 9 | Ca Na Cs Cl | −0.0030 | C | 0.00 | mp-989644 | $Na_1Cl_1 + 0.5\,Ca_2Cl_4 + 0.25\,Cl_4 + 2\,Cs_1Cl_1$ |
| 10 | Na In Cs Cl | −0.0033 | S | 2.99 | mp-989571 | $0.25\,Cs_6In_4Cl_{18} + 0.5\,Cs_1Cl_1 + Na_1Cl_1$ |
| 11 | Br S K Cl | −0.0044 | C | 0.00 | mp-989587 | $2\,K_1Cl_1 + 0.5\,Br_2Cl_2 + 0.25\,Cl_4 + 0.125\,S_8Cl_{16}$ |
| 12 | Tl In Cs H | −0.0052 | C | 0.00 | mp-996945 | $Tl_1 + 4.3333\,H_1 + 0.1111\,Cs_3In_9 + 1.6667\,Cs_1H_1$ |
| 13 | In Rb Cs Br | −0.0055 | S | 2.56 | mp-996941 | $Rb_1Br_1 + 0.5\,In_2Br_6 + 2\,Cs_1Br_1$ |
| 14 | I F Cs Cl | −0.0076 | S | 1.13 | mp-989516 | $0.1016\,Cl_{16} + 1.9375\,Cs_1Cl_1 + 0.0078\,Cs_8I_{24}F_{128} + 0.4062\,I_2Cl_6$ |
| 15 | Sr Na Cs F | −0.0090 | C | 0.00 | mp-989574 | $0.25\,F_4 + 2\,Cs_1F_1 + Sr_1F_2 + Na_1F_1$ |
| 16 | Sn Ca Cs Cl | −0.0092 | I | 3.61 | mp-989570 | $Cs_1Cl_1 + 0.25\,Cs_4Sn_4Cl_{12} + 0.5\,Ca_2Cl_4$ |
| 17 | Al In Rb H | −0.0100 | S | 0.14 | mp-989528 | $0.5\,Al_2H_6 + 1.75\,Rb_1H_1 + 0.25\,Rb_1In_4 + 1.25\,H_1$ |
| 18 | Tl Si Cs H | −0.0101 | C | 0.00 | mp-989560 | $0.25\,Si_4H_{16} + 2\,Cs_1H_1 + Tl_1$ |
| 19 | Br F K Cl | −0.0101 | S | 0.91 | mp-989591 | $2\,K_1Cl_1 + 0.5\,Br_2Cl_2 + 0.5\,Cl_4 + 0.25\,Cl_4F_4$ |
| 20 | N F Al Ca | −0.0103 | C | 0.00 | mp-989399 | $0.5\,Al_2N_2 + 0.25\,Ca_2Al_4 + 0.5\,Ca_1F_2 + 2.5\,Ca_2$ |
| 21 | Li In Cs Br | −0.0136 | S | 1.58 | mp-989405 | $2\,Cs_1Br_1 + 0.5\,In_2Br_6 + 0.5\,Li_2Br_2$ |
| 22 | Bi Na Cs Cl | −0.0138 | - | - | - | $0.1667\,Cs_{12}Bi_4Cl_{24} + Na_1Cl_1 + 0.0833\,Bi_4Cl_{12}$ |
| 23 | Sb Na Rb Cl | −0.0147 | I | 3.05 | mp-989545 | $0.25\,Sb_4Cl_{12} + 2\,Rb_1Cl_1 + Na_1Cl_1$ |
| 24 | N F In Ba | −0.0149 | C | 0.00 | mp-996942 | $0.5\,Ba_6N_2 + 0.5\,Ba_1F_2 + 0.5\,Ba_1 + 0.5\,Ba_4In_4$ |
| 25 | Tl Li Cs F | −0.0151 | I | 3.48 | mp-989562 | $0.6667\,Cs_3Tl_1F_6 + Li_1F_1 + 0.0833\,Tl_4F_{12}$ |
| 26 | Tl Ga Cs F | −0.0153 | I | 4.77 | mp-989558 | $0.0833\,Cs_{18}Ga_{12}F_{54} + 0.25\,Tl_4F_4 + 0.5\,Cs_1F_1$ |
| 27 | Cl Pb Cs F | −0.0167 | C | 0.01 | mp-989549 | $2\,Cs_1F_1 + 0.25\,Pb_4F_{12} + 0.25\,Cl_4F_4$ |
| 28 | S F Rb Cl | −0.0171 | C | 0.00 | mp-989388 | $0.1667\,S_1F_6 + 0.1042\,S_8Cl_{16} + 0.5833\,Cl_4 + 2\,Rb_1Cl_1$ |
| 29 | Ga Na Cs H | −0.0178 | S | 1.09 | mp-989594 | $0.5\,Na_2Ga_2H_8 + 2\,Cs_1H_1$ |
| 30 | In Li Tl F | −0.0187 | I | 4.02 | mp-989551 | $0.25\,Li_4In_4F_{16} + 0.5\,Tl_4F_4$ |
| 31 | Se N Pb Ca | −0.0197 | C | 0.00 | mp-989582 | $0.375\,Ca_8Pb_4 + Ca_1Se_1 + 0.25\,Ca_2Pb_2 + 0.0625\,Ca_{24}N_{16}$ |
| 32 | Tl In Rb F | −0.0207 | I | 3.66 | mp-989532 | $0.0833\,Rb_8In_{12}F_{44} + 0.25\,Tl_4F_4 + 1.3333\,Rb_1F_1$ |
| 33 | N F In Ca | −0.0208 | C | 0.00 | mp-989404 | $0.5\,Ca_8In_4N_2 + 0.5\,Ca_1F_2 + 0.75\,Ca_2$ |
| 34 | Pb Na Rb F | −0.0221 | C | 0.00 | mp-989569 | $2\,Rb_1F_1 + 0.25\,Pb_4F_{12} + Na_1F_1$ |
| 35 | In Bi Cs F | −0.0224 | S | 2.29 | mp-989538 | $0.0833\,Bi_4F_{12} + 0.5\,In_2F_6 + 0.3333\,Bi_2 + 2\,Cs_1F_1$ |
| 36 | Li Ga Tl F | −0.0226 | I | 4.26 | mp-989577 | $0.3333\,Ga_2F_6 + 0.5\,Tl_4F_4 + 0.0556\,Li_{18}Ga_6F_{36}$ |
| 37 | K In Rb Cl | −0.0227 | I | 3.54 | mp-989529 | $0.0833\,K_{12}In_4Cl_{24} + 2\,Rb_1Cl_1 + 0.3333\,In_2Cl_6$ |
| 38 | As Na Cs Cl | −0.0244 | I | 3.06 | mp-989608 | $0.25\,As_4Cl_{12} + 2\,Cs_1Cl_1 + Na_1Cl_1$ |
| 39 | Ca Na Cs F | −0.0246 | C | 0.00 | mp-989572 | $Na_1F_1 + Cs_2Ca_1F_4 + 0.25\,F_4$ |
| 40 | Li N Cs F | −0.0260 | S | 2.71 | mp-989536 | $0.75\,F_4 + 0.25\,N_4 + 0.25\,Cs_4Li_4F_8 + Cs_1F_1$ |
| 41 | Tl Na Rb F | −0.0262 | I | 3.64 | mp-989548 | $0.25\,Rb_4Tl_4F_{16} + Rb_1F_1 + Na_1F_1$ |
| 42 | Ga Rb Cs F | −0.0278 | I | 5.94 | mp-989618 | $0.0833\,Cs_{18}Ga_{12}F_{54} + Rb_1F_1 + 0.5\,Cs_1F_1$ |

*Continued on next page*

Table 4.3 – *Continued from previous page*

| # | Elpasolite $x_1$ | $x_2$ | $x_3$ | $x_4$ | $\Delta E$ (eV/atom) | Type | $\Delta\epsilon$ (eV) | MP ID | Competing mixed phase in MP |
|---|---|---|---|---|---|---|---|---|---|
| 43 | Tl | In | Cs | Br | −0.0282 | S | 1.83 | mp-996944 | $Tl_1Br_1 + 0.5\,In_2Br_6 + 2\,Cs_1Br_1$ |
| 44 | Na | In | Cs | Br | −0.0285 | S | 1.89 | mp-996943 | $Na_1Br_1 + 0.5\,In_2Br_6 + 2\,Cs_1Br_1$ |
| 45 | In | Ga | Tl | F | −0.0295 | S | 2.36 | mp-989555 | $0.6667\,In_1 + 0.1667\,In_2F_6 + 0.5\,Tl_4F_4 + 0.5\,Ga_2F_6$ |
| 46 | Tl | Ga | Cs | H | −0.0298 | S | 0.38 | mp-989553 | $4.3333\,H_1 + 0.1111\,Cs_3Ga_9 + 1.6667\,Cs_1H_1 + Tl_1$ |
| 47 | Rb | Tl | Cs | F | −0.0299 | I | 3.74 | mp-989567 | $0.0833\,Rb_4Tl_4F_{16} + 0.6667\,Rb_1F_1 + 0.6667\,Cs_3Tl_1F_6$ |
| 48 | In | Li | Tl | Cl | −0.0306 | S | 2.87 | mp-989512 | $0.0556\,Li_{18}In_6Cl_{36} + 2\,Tl_1Cl_1 + 0.3333\,In_2Cl_6$ |
| 49 | Ga | Al | Cs | H | −0.0309 | S | 0.86 | mp-989648 | $0.5\,Al_2H_6 + 1.3333\,H_1 + 1.6667\,Cs_1H_1 + 0.1111\,Cs_3Ga_9$ |
| 50 | Cl | N | Tl | Ba | −0.0312 | C | 0.00 | mp-989542 | $0.5\,Ba_6N_2 + 0.5\,Ba_1Cl_2 + 0.5\,Ba_2Tl_4 + 1.5\,Ba_1$ |
| 51 | Tl | K | Cs | F | −0.0329 | I | 3.82 | mp-989526 | $0.6667\,Cs_3Tl_1F_6 + 0.0417\,K_{24}Tl_8F_{48}$ |
| 52 | N | O | Sn | Sr | −0.0333 | C | 0.00 | mp-989540 | $Sr_2N_1 + Sr_3Sn_1O_1 + 0.5\,Sr_2Sn_2$ |
| 53 | S | F | Cs | Cl | −0.0336 | C | 0.00 | mp-989521 | $0.1458\,Cl_{16} + 0.1667\,S_1F_6 + 0.1042\,S_8Cl_{16} + 2\,Cs_1Cl_1$ |
| 54 | Se | Cl | Cs | F | −0.0349 | C | 0.03 | mp-989544 | $0.5\,Cs_1Cl_1 + 0.125\,Cl_4F_4 + 1.5\,Cs_1F_1 + 0.25\,Se_4F_{16}$ |
| 55 | Ga | K | Cs | F | −0.0359 | I | 6.04 | mp-989531 | $0.5\,Cs_1F_1 + 0.0833\,Cs_{18}Ga_{12}F_{54} + K_1F_1$ |
| 56 | Pb | Rb | Cs | F | −0.0370 | C | 0.00 | mp-989525 | $2\,Cs_1F_1 + 0.25\,Pb_4F_{12} + Rb_1F_1$ |
| 57 | F | Br | Rb | Cl | −0.0382 | S | 0.92 | mp-989573 | $0.5\,Cl_4 + 0.5\,Br_2Cl_2 + 2\,Rb_1Cl_1 + 0.25\,Cl_4F_4$ |
| 58 | N | Rb | Cs | F | −0.0384 | S | 2.96 | mp-989519 | $0.5\,Rb_2F_6 + 0.125\,N_8 + 2\,Cs_1F_1 + 0.25\,F_4$ |
| 59 | N | Li | Na | F | −0.0396 | S | 2.72 | mp-989504 | $0.375\,F_8 + 0.5\,N_2 + Li_1F_1 + 2\,Na_1F_1$ |
| 60 | Bi | Na | Rb | Cl | −0.0416 | I | 3.73 | mp-989520 | $0.25\,Bi_4Cl_{12} + 2\,Rb_1Cl_1 + Na_1Cl_1$ |
| 61 | Na | Mg | Cs | F | −0.0449 | C | 0.00 | mp-989568 | $0.25\,F_4 + 0.6667\,Cs_1F_1 + 0.1667\,Cs_8Mg_6F_{20} + Na_1F_1$ |
| 62 | Tl | Al | Rb | H | −0.0451 | S | 0.72 | mp-989539 | $H_1 + 0.5\,Al_2H_6 + Tl_1 + 2\,Rb_1H_1$ |
| 63 | S | Br | Rb | Cl | −0.0452 | C | 0.00 | mp-989518 | $0.25\,Cl_4 + 0.5\,Br_2Cl_2 + 2\,Rb_1Cl_1 + 0.125\,S_8Cl_{16}$ |
| 64 | N | F | Sn | Sr | −0.0454 | C | 0.00 | mp-989592 | $0.5\,Sr_8Sn_4 + 0.0625\,Sr_{32}N_{16}F_{16}$ |
| 65 | As | Na | Rb | F | −0.0467 | I | 4.55 | mp-989523 | $2\,Rb_1F_1 + 0.25\,As_4F_{12} + Na_1F_1$ |
| 66 | Pb | K | Cs | F | −0.0486 | C | 0.00 | mp-989585 | $2\,Cs_1F_1 + 0.25\,Pb_4F_{12} + K_1F_1$ |
| 67 | In | Al | Cs | H | −0.0488 | S | 0.61 | mp-989535 | $0.5\,Al_2H_6 + 0.1111\,Cs_3In_9 + 1.6667\,Cs_1H_1 + 1.3333\,H_1$ |
| 68 | Pb | Na | Cs | F | −0.0513 | C | 0.00 | mp-989556 | $2\,Cs_1F_1 + 0.25\,Pb_4F_{12} + Na_1F_1$ |
| 69 | Ga | Na | Tl | F | −0.0514 | I | 4.39 | mp-989561 | $0.2\,Ga_2F_6 + 0.5\,Tl_4F_4 + 0.1\,Na_{10}Ga_6F_{28}$ |
| 70 | In | Tl | Rb | Cl | −0.0528 | S | 2.40 | mp-989550 | $0.5\,In_2Cl_6 + Tl_1Cl_1 + 2\,Rb_1Cl_1$ |
| 71 | In | Na | Rb | Cl | −0.0538 | I | 3.05 | mp-989547 | $0.3333\,In_2Cl_6 + 0.1667\,Na_6In_2Cl_{12} + 2\,Rb_1Cl_1$ |
| 72 | In | Na | Cs | H | −0.0565 | S | 1.26 | mp-989610 | $3.3333\,H_1 + 0.1111\,Cs_3In_9 + 1.6667\,Cs_1H_1 + Na_1H_1$ |
| 73 | Li | In | Rb | Cl | −0.0603 | S | 2.83 | mp-989583 | $0.0556\,Li_{18}In_6Cl_{36} + 2\,Rb_1Cl_1 + 0.3333\,In_2Cl_6$ |
| 74 | In | Na | Tl | F | −0.0604 | I | 4.27 | mp-989533 | $Na_1F_1 + 0.5\,Tl_4F_4 + 0.5\,In_2F_6$ |
| 75 | Br | F | Cs | Cl | −0.0610 | S | 0.95 | mp-989543 | $0.5\,Cl_4 + 2\,Cs_1Cl_1 + 0.5\,Br_2Cl_2 + 0.25\,Cl_4F_4$ |
| 76 | In | Ga | Rb | F | −0.0616 | I | 3.27 | mp-989566 | $0.1667\,Ga_4 + 0.0833\,Rb_8In_{12}F_{44} + 0.1667\,Ga_2F_6 + 1.3333\,Rb_1F_1$ |
| 77 | N | K | Cs | F | −0.0618 | I | - | mp-989580 | $0.25\,F_4 + 0.125\,K_8F_{24} + 0.0625\,N_{16} + 2\,Cs_1F_1$ |
| 78 | Li | Na | Cs | F | −0.0634 | C | 0.00 | mp-989559 | $0.5\,F_4 + Cs_1F_1 + 0.25\,Cs_4Li_4F_8 + Na_1F_1$ |
| 79 | Na | In | Rb | F | −0.0672 | I | 5.34 | mp-989578 | $0.0833\,Rb_8In_{12}F_{44} + Na_1F_1 + 1.3333\,Rb_1F_1$ |
| 80 | O | N | Sn | Ca | −0.0686 | C | 0.00 | mp-989584 | $0.0625\,Ca_{24}N_{16} + Ca_3Sn_1O_1 + 0.0227\,Ca_{62}Sn_{40} + 0.0455\,Ca_2Sn_2$ |
| 81 | N | F | Sn | Ca | −0.0738 | S | 0.14 | mp-989590 | $0.5\,Ca_4N_2F_2 + 0.5\,Ca_8Sn_4$ |
| 82 | In | Rb | Cs | F | −0.0776 | I | 5.37 | mp-989605 | $0.0833\,Rb_8In_{12}F_{44} + 0.3333\,Rb_1F_1 + 2\,Cs_1F_1$ |
| 83 | S | Br | Cs | Cl | −0.0848 | C | 0.00 | mp-989517 | $2\,Cs_1Cl_1 + 0.5\,Br_2Cl_2 + 0.25\,Cl_4 + 0.125\,S_8Cl_{16}$ |
| 84 | In | K | Cs | F | −0.0875 | I | 5.46 | mp-989639 | $0.025\,K_{24}In_8F_{48} + 0.1\,K_4In_8F_{28} + 2\,Cs_1F_1$ |
| 85 | Tl | Al | Cs | H | −0.0884 | S | 1.14 | mp-989575 | $2\,Cs_1H_1 + 0.5\,Al_2H_6 + Tl_1 + H_1$ |
| 86 | Tl | Ga | Rb | F | −0.0945 | I | 4.40 | mp-989565 | $2\,Rb_1F_1 + 0.5\,Ga_2F_6 + 0.25\,Tl_4F_4$ |
| 87 | Ga | Na | Rb | F | −0.1008 | I | 5.90 | mp-989400 | $2\,Rb_1F_1 + 0.1\,Na_{10}Ga_6F_{28} + 0.2\,Ga_2F_6$ |
| 88 | Al | Na | Cs | H | −0.1019 | S | 2.14 | mp-989642 | $2\,Cs_1H_1 + 0.5\,Na_2Al_2H_8$ |
| 89 | N | Na | Cs | F | −0.1064 | S | 2.80 | mp-989527 | $0.75\,F_4 + 0.5\,N_2 + Na_1F_1 + 2\,Cs_1F_1$ |
| 90 | Tl | In | Cs | F | −0.1092 | I | 3.99 | mp-989537 | $2\,Cs_1F_1 + 0.25\,Tl_4F_4 + 0.5\,In_2F_6$ |

Among these Elpasolites, metals, semiconductors and insulators are roughly distributed equally. All structures with an earth alkaline metal in crystal position 4 have a low or zero band-gap. We have noted an intriguing yet stable structure of a conductor, $NFAl_2Ca_6$ (MP ID: mp-989399, # 20 in Table 4.2) with Ca at position 4, instead of F or Cl.

Due to the unusual composition and atomic charges of $NFAl_2Ca_6$ (MP ID: mp-989399), a more detailed analysis of its stability has been carried out. More specifically, we investigated if any

common competing phases which are amiss in the MP would push $NFAl_2Ca_6$ off the convex hull. $NF_3$ is the only compound found to be missing from MPs data base. The formation energy of $NF_3$ were calculated using a cubic cell with fixed lattice vectors of length 10Å and a 1x1x1 Monkhorst-Pack $k$-mesh. The calculated formation energy of $NF_3$ (-0.465 eV/atom) is insufficient to push $NFAl_2Ca_6$ off the convex hull and would need a formation energy of ∼-4.0 eV/atom in order do so.

We also perturbed the structure in order to see if it relaxes back into its original geometry. Three perturbed structures were generated by randomly dislocating all atoms in the cell, with a mean absolute deviation of ∼0.2Å from the original structure. For one of the structures we also changed lattice parameters ($a$, $b$, $c$, $\alpha$, $\beta$ and $\gamma$ from 6.96, 6.96, 6.96, 60.0, 60.0 and 60.0 to 7.96, 6.96, 6.96, 60.0, 85.0 and 60.0, respectively) and observed relaxation. Initial structures have formation energies ∼1 eV/atom higher than the original structure, and relax back into the original configuration.

Finally, phonon spectra were calculated along high symmetry lines in the Brillouin zone, using phonopy [217] with a 2x2x2 super-cell. The spectra, see Figure 4.13, revealed neither negative nor imaginary phonon frequencies.

Bader charge analysis [218–220] (Table 4.4) indicates an exotic negative oxidation state for Al (-II), previously only reported for Al in substantially larger Zintl phase unit cells ($Sr_{14}[Al_4]_2Ge_3$) [221]. Since Bader charges sometimes yield non intuitive results [222, 223], calculated Hirshfeld [224] and Voronoi deformation density [222, 225] charges (Table 4.4) confirm the negative oxidation state, albeit reduced by one unit (-I). The calculations used for the Hirshfeld and Voronoi deformation densities were done using SIESTA[226]. These calculations were done using a triple-$\zeta$ basis with double polarization radial functions. An energy shift of 0.01 eV to control the diffusion of the atomic orbitals, a real-space energy cutoff of 360 Ry for the charge density, and a 6x6x6 Monkhorst-Pack $k$-mesh were used. Electronic temperature was set to 0.1 eV and the PBE functionals. The pseudo potentials of F, N, and Al were extracted from SIESTAs pseudo potential library. Semi core $3p$-states were included in the valence for Ca. The cutoff radius used for each angular component to generate pseudo potentials were chosen to be 2.8, 1.4, 1.6, and 1.9 Bohr for $4s$, $3p$, $3d$ and $4f$, respectively.

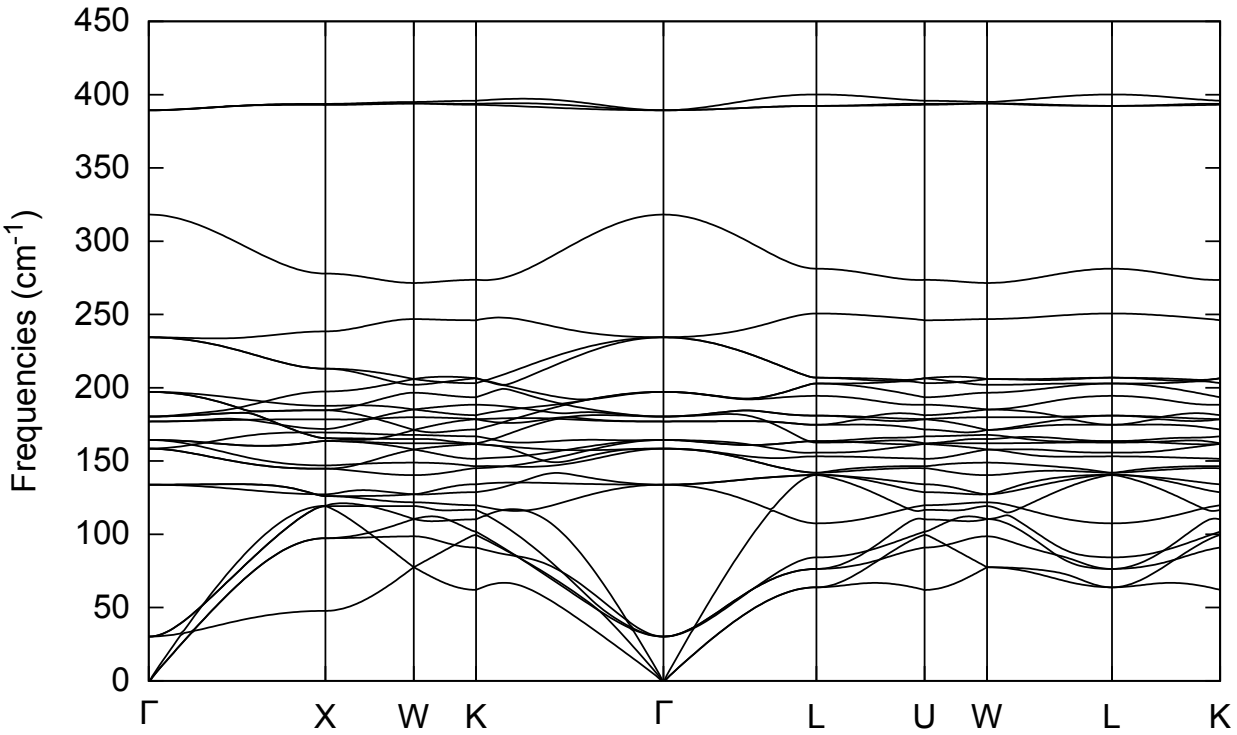



**Figure 4.13:** Calculated phononspectra along high symetry lines of $NFAl_2Ca_6$ (MP ID: mp-989399). Vertical axis is the phonon strength, given in $cm^{-1}$. Horizontal axis depicts traversment inbetween symetry points. The figure reveals that the structure has no negative or imaginary phonons.

**Table 4.4:** Calculated atomic charges in $NFAl_2Ca_6$ Elpasolite using different methods (obtained using `SIESTA`[226]).

| Method | N | F | Al | Ca |
|---|---|---|---|---|
| Bader | −2.00 | −0.98 | −2.13 | 1.20 |
| Hirshfeld | −0.63 | −0.36 | −1.05 | 0.52 |
| Voronoi deformation density | −0.81 | −0.29 | −1.13 | 0.56 |

## 7.5 Conclusion

In conclusion, we have developed and used ML-models of formation energies to investigate all possible Elpasolites made up of main-group elements. We have presented numerical results for ∼2 M formation energies. The ML-model is only implicitly dependent on spatial coordinates, through reference data used for training. No spatial coordinates are needed for new queries, yet for a training set of 10 k crystals the model reaches ±0.1 eV/atom—comparable to DFT accuracy for solids. The results have been used to identify the most strongly bound Elpasolites as well as to investigate energy and bonding trends at crystal structure sites, leading to a new "Elpasolite order" of elements, consistent with the bonding physics in the Elpasolite crystal structure.

We identified and added 128 structures (90 unique stoichiometries) to the convex hull of the MP database. Charge analysis for the metallic Elpasolite $NFAl_2Ca_6$ indicates a negative atomic oxidation state of Al. This outcome directly demonstrates that our method can be used for the discovery of stable as well as unconventional chemistries. Due to the low computational cost of the ML model one can now also afford to remove human bias by considering also those structures which previously would have been excluded due to "chemical intuition". Our results suggest that ML models hold great promise for the computational screening of polymorphs, other crystal structure symmetries, solid mixtures, phase transitions, or defects at unprecedented rate and extent. Other crystal properties than energies could also be considered.

# Chapter 8

# Concluding Remarks

To summarize, this thesis has been dedicated to developing, benchmarking, and exploring the applicability of QML models.

A significant portion has been dedicated to benchmarking the performance of numerous combinations of regressors and representations for several electronic ground-state properties from the QM9 data set. The results indicate that for all properties, at least one QML model is capable of surpassing the accuracy threshold the B3LYP DFT generated training data with respect to experimental data.

Furthermore, at least one QML model is capable of producing out-of-sample errors on par with chemical accuracy, or better, as is the case for out of 7 out of 12 distinct properties (atomization energies, heat-capacity, $\omega_1$, $\mu$). For the remaining properties $\alpha$, $\varepsilon_{\text{HOMO}}$, $\varepsilon_{\text{LUMO}}$, $\Delta\epsilon$, and ZPVE, the errors of the best models are within two factors of chemical accuracy.

While these results do not guarantee that QML models will perform comparably when trained on data calculated across higher levels of theory, prior studies do provide evidence to this effect [31].

Therefore, in order to further reduce the QML model's predictive error with respect to experimental values, one would have to improve the quality of the training data itself.

Another area of focus has been the development of new and accurate QML models. For example, a new representation for quantum machine learning that uses a sum of multidimensional Gaussians placed on elemental, atom-pairwise, and angular degrees of freedom to represent the atomic chemical environment has been developed.

Numerical results demonstrate that, compared to current benchmarks, QML models using this representation show superior predictive power on a diverse set of systems, including diverse organic molecules, non-covalently bonded protein side-chains, water clusters, and crystalline

solids. Furthermore, the QML model produces semi-qualitative covalent bonding potentials for single, double, and triple bonds containing chemical elements withheld during training.

The architecture of a machine learning model has also been proven to be a critical component of the model's accuracy. For example, the use of corresponding response operators as proxies for learning energy response properties leads to QML models with significantly lower out-of-sample errors than learning the corresponding properties directly. The formalism can also be used to reproduce accurate molecular normal modes and IR-spectra.

Finally, the thesis explores the applicability of QML models. A QML model was used to learn the formation energies of all $\sim$2 M possible elpasolite crystal structures comprised of main-group elements. The model encoded spatial information implicitly, and did not need specific coordinates for new queries while reaching a mean absolute error of $\pm$0.1 eV/atom when trained on only 10 k crystals. The resulting energy predictions were used to investigate energy and bonding trends at crystal structure sites. Furthermore, 90 stoichiometries were identified to lie on the convex hull of stability, and charge analysis indicated that one of the structures, $NFAl_2Ca_6$, possessed a negative atomic oxidation state of Al. This outcome directly demonstrated that our method could be used for the discovery of stable as well as unconventional chemistries.

Due to the low computational cost of the ML model, one can now also afford to remove human bias by also considering those structures which previously would have been excluded due to "chemical intuition".

This thesis is fittingly concluded with some general observations and remarks on QML.

Quantum chemistry and material science are undergoing a paradigm shift, where machine learning models will complement and improve upon a large portion of existing computational methods and speed up the process of discovering, amongst other things, new drugs and materials.

The combination of this vast quantity of data and machine learning holds great promise for future research and for the discovery of new and exotic compounds.

The field is, however, still in its infancy, and there are many challenges and problems to address, including finding optimal ways of selecting training data, how to best represent a compound in a machine learning model, and how to best learn electronic excitations.

# Appendix A

# Derivation of Fourier series used for angular binning

This section is dedicated to derive the Fourier coefficients $c_n$ and $s_n$ we use to speed up the $A_3$ and $A_4$ scalar products.

First, let $f_T(x)$ be a function consisting of $2N$ functions $g(x)$ placed at a periodically with period $T$ and a shift $\theta \in [0, T]$, as in eq. 0.1.

$$f_T(x) = \frac{1}{N} \sum_{j=-N}^{N} g(x - \theta - jT) \tag{0.1}$$

The Fourier transform of $f_T(x)$ will then be:

$$\begin{aligned} \widehat{f_T}(\omega) &= \frac{1}{N} \sum_{j=-N}^{N} \hat{g}(\omega) \exp[-i\omega(\theta - jT)] \\ &= \frac{1}{N} \hat{g}(\omega) \exp[-i\omega\theta] \left[ 1 + 2 \sum_{j=1}^{N} \cos[\omega(jT)] \right] \\ &= \frac{1}{N} \hat{g}(\omega) \exp[-i\omega\theta] \left[ 2 \frac{\sin[\frac{(N+1)\omega T}{2}] \cos[\omega(\frac{NT}{2})]}{\sin[\frac{T\omega}{2}]} - 1 \right] \end{aligned} \tag{0.2}$$

Now we let $N$ tend to infinity. $\widehat{f_T}(\omega) \to 0$ for $\omega \neq \frac{2\pi n}{T}$ when $N \to \infty$, since $\frac{\sin[\frac{(N+1)\omega T}{2}] \cos[\omega(\frac{NT}{2})]}{\sin[\frac{T\omega}{2}]}$ is bounded and $\frac{1}{N} \to 0$. If $\omega \to \frac{2\pi n}{T}$, then:

$$\lim_{\omega\to\frac{2\pi n}{T}} \frac{\sin[\frac{(N+1)\omega T}{2}]}{\sin[\frac{T\omega}{2}]} = (1+N)\frac{\cos(n(1+N)\pi)}{\cos(n\pi)} \Rightarrow$$

$$\lim_{\omega\to\frac{2\pi n}{T}} \widehat{f_T}(\omega) = \frac{1}{N}\widehat{g}(\frac{2\pi n}{T})\exp[-i\frac{2\pi\theta n}{T}]\left[(1+N)\frac{\cos(n(1+N)\pi)}{\cos(n\pi)}\cos[\pi nN]-1\right] \Rightarrow$$

$$\Rightarrow \Big/ \frac{\cos(n(1+N)\pi)}{\cos(n\pi)}\cos[\pi nN] = 1 \Big/ \Rightarrow$$

$$\lim_{\omega\to\frac{2\pi n}{T}} \widehat{f_T}(\omega) = \widehat{g}(\frac{2\pi n}{T})\exp[-i\frac{2\pi\theta n}{T}]$$

We now define $a_n$ as the limit:

$$a_n \equiv \lim_{\omega\to n} \widehat{f_{2\pi}}(\omega)$$

which is equal to the Fourier coefficients:

$$a_n = \widehat{g}(n)\exp(-in\theta)$$

or the cosine and sine coefficients.

$$\begin{aligned} c_n = a_n + a_{-n} = &\quad \widehat{g}(n)\cos(n\theta), n>0 \\ s_n = \frac{a_n - a_{-n}}{i} = &\quad \widehat{g}(n)\sin(n\theta), n>0 \\ a_0 = &\quad \widehat{g}(0) \end{aligned}$$

Letting $f_{2\pi}(x) = \Theta(x,\sigma)$ leads to the following Fourier coefficients:

$$\begin{aligned} g(x) &= \exp[-\frac{(x-\theta)^2}{2\sigma^2}] - \exp[-\frac{(x-\theta-\pi)^2}{2\sigma^2}] \\ a_n &= \sigma\sqrt{2\pi}\exp[-\frac{(\sigma n)^2}{2}](\exp[-in\theta] - \exp[-in(\pi+\theta)]) \\ c_n &= \sigma\sqrt{8\pi}\exp[-\frac{(\sigma n)^2}{2}](\cos[\theta n] - \cos[(\pi+\theta)n]), n>0 \\ s_n &= \sigma\sqrt{8\pi}\exp[-\frac{(\sigma n)^2}{2}](\sin[\theta n] - \sin[(\pi+\theta)n]), n>0 \\ a_0 &= 0 \end{aligned} \tag{0.3}$$

$$||a|| \equiv \sqrt{\sum_{n=-\infty}^{\infty} a_n \bar{a}_n} \tag{0.4}$$
$$a_n \bar{a}_n = \sigma^2 2\pi \exp[-(\sigma n)^2](2 - 2\cos[n\pi])$$

# Appendix B

# Derivation of Operators

The total potential energy $U_C^*$ of a query compound $C$ can be decomposed into a sum of local energy contributions which are calculated using a weighted sum of kernels, seen in Eq. 0.1. The sum runs over all atomic environments $I$ in in the query compound.

$$U_C^* = \sum_{I \in C} U_{\text{local}}^* \left(q_I^*\right) = \sum_{I \in C} \sum_J \mathcal{k}\left(q_J, q_I^*\right) \alpha_J \tag{0.1}$$

A response property, $\omega$, corresponding to the response operator, $\mathcal{O}$ acting on the energy, $U$, can then be calculated as:

$$\omega = \mathcal{O}[\mathbf{U}] \approx \mathcal{O}[\mathbf{K}]\boldsymbol{\alpha} \tag{0.2}$$

The optimal set of regression coefficients, $\boldsymbol{\alpha}$ can be obtained by minimizing the following Lagrangian.

$$\begin{aligned}
J(\boldsymbol{\alpha}) =& \sum_\gamma \beta_\gamma \|\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}}) - \mathcal{O}_\gamma(K\boldsymbol{\alpha})\|^2_{L_2(\Omega_\gamma)} \\
\equiv& \sum_\gamma \beta_\gamma \int_{\Omega_\gamma} [\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}}) - \mathcal{O}_\gamma(K\boldsymbol{\alpha})]^T [\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}}) - \mathcal{O}_\gamma(K\boldsymbol{\alpha})] \\
=& \sum_\gamma \beta_\gamma \int_{\Omega_\gamma} [\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}})]^T [\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}})] + [\mathcal{O}_\gamma(K\boldsymbol{\alpha})]^T [\mathcal{O}_\gamma(K\boldsymbol{\alpha})] \\
&-2[\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}})]^T [\mathcal{O}_\gamma(K\boldsymbol{\alpha})] \\
=& \sum_\gamma \beta_\gamma \int_{\Omega_\gamma} [\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}})]^T [\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}})] + \boldsymbol{\alpha}^T [\mathcal{O}_\gamma(K)]^T [\mathcal{O}_\gamma(K)]\boldsymbol{\alpha} \\
&-2[\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}})]^T [\mathcal{O}_\gamma(K)]\boldsymbol{\alpha}
\end{aligned}$$

We define the integral over the integration manifold as 1:

$$\int_{\Omega_\gamma} = 1 \tag{0.3}$$

The derivative of the Lagrangian is given by:

$$\begin{aligned}
\frac{dJ(\mathbf{U}^{\text{ref}}, \boldsymbol{\alpha})}{d\boldsymbol{\alpha}} =& \sum_\gamma \beta_\gamma \int_{\Omega_\gamma} [\mathcal{O}_\gamma(K)]^T [\mathcal{O}_\gamma(K)]\boldsymbol{\alpha} + \boldsymbol{\alpha}^T [\mathcal{O}_\gamma(K)]^T [\mathcal{O}_\gamma(K)] \\
&-2[\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}})]^T [\mathcal{O}_\gamma(K)] \\
=& \sum_\gamma \beta_\gamma \int_{\Omega_\gamma} 2[\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}})]^T [\mathcal{O}_\gamma(K)]\boldsymbol{\alpha} - 2[\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}})]^T [\mathcal{O}_\gamma(K)] \\
=& 2\sum_\gamma \beta_\gamma \Big( \int_{\Omega_\gamma} [\mathcal{O}_\gamma(K)]^T [\mathcal{O}_\gamma(K)]\boldsymbol{\alpha} \\
&- \int_{\Omega_\gamma} 2[\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}})]^T [\mathcal{O}_\gamma(K)] \Big) \\
=& 2\Big( \sum_\gamma \beta_\gamma \int_{\Omega_\gamma} [\mathcal{O}_\gamma(K)]^T [\mathcal{O}_\gamma(K)]\boldsymbol{\alpha} \Big) \\
&-2\Big( \sum_\gamma \beta_\gamma \int_{\Omega_\gamma} 2[\mathcal{O}_\gamma(U)]^T [\mathcal{O}_\gamma(K)] \Big)
\end{aligned}$$

We now arrive at the corresponding normal-equation solution to the problem:

$$0 = \frac{dJ(\boldsymbol{\alpha})}{d\boldsymbol{\alpha}} \Leftrightarrow \tag{0.4}$$

$$\boldsymbol{\alpha} = \Big( \sum_\gamma \beta_\gamma \int_{\Omega_\gamma} [\mathcal{O}_\gamma(K)]^T [\mathcal{O}_\gamma(K)] \Big)^{-1} \Big( \sum_\gamma \beta_\gamma \int_{\Omega_\gamma} [\mathcal{O}_\gamma(\mathbf{U}^{\text{ref}})]^T [\mathcal{O}_\gamma(K)] \Big) \tag{0.5}$$

**First-Order Differential Operators**

Many response operators in chemistry correspond to the gradient of the energy with respect a change in 3-dimensional variable, $\eta$, such as the nuclear coordinates or an externally applied magnetic or electric field. Here we show the solution to any first-order 3D differential operator acting on the energy and kernel.

The domain of integration, $\Omega$, is the gradient projected on a sphere.

$$\Omega = \{\eta_x, \eta_y, \eta_z \in \mathbb{R} | \eta_x^2 + \eta_x^2 + \eta_x^2 = (4\pi)^{-1}\} \tag{0.6}$$

First we project the gradient on a spherical coordinate basis, $\mathbf{r}$:

$$\mathcal{O} = \nabla_\eta^{\theta\phi} = \nabla_\eta \cdot \mathbf{r} \tag{0.7}$$

Where the gradient is given by:

$$\nabla_\eta = (\frac{\partial}{\partial \eta_x}, \frac{\partial}{\partial \eta_y}, \frac{\partial}{\partial \eta_z}) \tag{0.8}$$

and the normal vector of the sphere on which the gradient is projected.

$$\mathbf{r}(\phi, \theta) = (4\pi)^{-\frac{1}{2}}(\cos(\phi)\sin(\theta), \sin(\phi)\sin(\theta), \cos(\theta)) \tag{0.9}$$

The integrals in the Lagrangian (Eq. 0.5) which corresponds to integrating out rotational degrees of freedom are for the left-hand side

$$\begin{aligned}
&\int_\Omega [\mathcal{O}(K)]^T[\mathcal{O}(K)] = \int_0^\pi \int_0^{2\pi} [\nabla_\eta^{\theta\phi} K]^T [\nabla_\eta^{\theta\phi} K] \sin\theta d\theta d\phi \\
=&\int_0^\pi \int_0^{2\pi} \Big[\frac{\partial K}{\partial \eta_x}\cos(\phi)\sin(\theta) + \frac{\partial K}{\partial \eta_y}\sin(\phi)\sin(\theta) + \frac{\partial K}{\partial \eta_z}\cos(\theta)\Big]^T \\
&\Big[\frac{\partial K}{\partial \eta_x}\cos(\phi)\sin(\theta) + \frac{\partial K}{\partial \eta_y}\sin(\phi)\sin(\theta) + \frac{\partial K}{\partial \eta_z}\cos(\theta)\Big] \sin\theta d\theta d\phi \\
=&\frac{1}{4\pi}\int_0^\pi \int_0^{2\pi} \Big(\frac{\partial K}{\partial \eta_x}^T \frac{\partial K}{\partial \eta_x}\cos^2(\phi)\sin^2(\theta) + \frac{\partial K}{\partial \eta_y}^T \frac{\partial K}{\partial \eta_y}\sin^2(\phi)\sin^2(\theta) + \frac{\partial K}{\partial \eta_z}^T \frac{\partial K}{\partial \eta_z}\cos^2(\theta) \\
&+(\frac{\partial K}{\partial \eta_x}^T \frac{\partial K}{\partial \eta_y} + \frac{\partial K}{\partial \eta_y}^T \frac{\partial K}{\partial \eta_x})\cos(\phi)\sin(\phi)\sin^2(\theta) \\
&+(\frac{\partial K}{\partial \eta_x}^T \frac{\partial K}{\partial \eta_z} + \frac{\partial K}{\partial \eta_z}^T \frac{\partial K}{\partial \eta_x})\cos(\phi)\sin(\theta)\cos(\theta) \\
&+(\frac{\partial K}{\partial \eta_z}^T \frac{\partial K}{\partial \eta_y} + \frac{\partial K}{\partial \eta_y}^T \frac{\partial K}{\partial \eta_z})\sin(\phi)\sin(\theta)\cos(\theta)\Big) \sin\theta d\theta d\phi \\
=&\frac{1}{3}\Big(\frac{\partial K}{\partial \eta_x}^T \frac{\partial K}{\partial \eta_x} + \frac{\partial K}{\partial \eta_y}^T \frac{\partial K}{\partial \eta_y} + \frac{\partial K}{\partial \eta_z}^T \frac{\partial K}{\partial \eta_z}\Big)
\end{aligned}$$

and the right-hand side

$$
\begin{aligned}
&\int_\Omega [\mathcal{O}(U)]^T[\mathcal{O}(K)] = \int_0^\pi \int_0^{2\pi} [\nabla_\eta^{\theta\phi} U]^T [\nabla_\eta^{\theta\phi} K] \sin\theta d\theta d\phi \\
=&\int_0^\pi \int_0^{2\pi} \Big[\frac{\partial U}{\partial \eta_x}\cos(\phi)\sin(\theta) + \frac{\partial U}{\partial \eta_y}\sin(\phi)\sin(\theta) + \frac{\partial U}{\partial \eta_z}\cos(\theta)\Big]^T \\
&\Big[\frac{\partial K}{\partial \eta_x}\cos(\phi)\sin(\theta) + \frac{\partial K}{\partial \eta_y}\sin(\phi)\sin(\theta) + \frac{\partial K}{\partial \eta_z}\cos(\theta)\Big] \sin\theta d\theta d\phi \\
=&\frac{1}{4\pi}\int_0^\pi \int_0^{2\pi} \Big( \frac{\partial U}{\partial \eta_x}^T \frac{\partial K}{\partial \eta_x}\cos^2(\phi)\sin^2(\theta) + \frac{\partial U}{\partial \eta_y}^T \frac{\partial K}{\partial \eta_y}\sin^2(\phi)\sin^2(\theta) + \frac{\partial U}{\partial \eta_z}^T \frac{\partial K}{\partial \eta_z}\cos^2(\theta) \\
&+(\frac{\partial U}{\partial \eta_x}^T \frac{\partial K}{\partial \eta_y} + \frac{\partial U}{\partial \eta_y}^T \frac{\partial K}{\partial \eta_x})\cos(\phi)\sin(\phi)\sin^2(\theta) \\
&+(\frac{\partial U}{\partial \eta_x}^T \frac{\partial K}{\partial \eta_z} + \frac{\partial U}{\partial \eta_z}^T \frac{\partial K}{\partial \eta_x})\cos(\phi)\sin(\theta)\cos(\theta) \\
&+(\frac{\partial U}{\partial \eta_z}^T \frac{\partial K}{\partial \eta_y} + \frac{\partial U}{\partial \eta_y}^T \frac{\partial K}{\partial \eta_z})\sin(\phi)\sin(\theta)\cos(\theta)\Big) \sin\theta d\theta d\phi \\
=&\frac{1}{3}\Big( \frac{\partial U}{\partial \eta_x}^T \frac{\partial K}{\partial \eta_x} + \frac{\partial U}{\partial \eta_y}^T \frac{\partial K}{\partial \eta_y} + \frac{\partial U}{\partial \eta_z}^T \frac{\partial K}{\partial \eta_z}\Big)
\end{aligned}
$$

**Second-Order Differential Operators**

Similarly, we define a second-order differential operator, $\mathcal{H}$, e.g. the Hessian of the energy with respect to the nuclear coordinates:

$$
\mathcal{H} = \nabla_{\eta_1} \otimes \nabla_{\eta_2} = \begin{bmatrix} \frac{\partial^2}{\partial\eta_x\partial\eta'_x} & \frac{\partial^2}{\partial\eta_x\partial\eta'_y} & \frac{\partial^2}{\partial\eta_x\partial\eta'_z} \\ \frac{\partial^2}{\partial\eta_y\partial\eta'_x} & \frac{\partial^2}{\partial\eta_y\partial\eta'_y} & \frac{\partial^2}{\partial\eta_y\partial\eta'_z} \\ \frac{\partial^2}{\partial\eta_z\partial\eta'_x} & \frac{\partial^2}{\partial\eta_z\partial\eta'_y} & \frac{\partial^2}{\partial\eta_z\partial\eta'_z} \end{bmatrix} \tag{0.10}
$$

We project the operator on the spherical coordinate basis:

$$
\mathcal{O} = \mathcal{H}^{\theta\phi\theta'\phi'} = \mathbf{r}\cdot\mathcal{H}\cdot\mathbf{r}' = \mathbf{r}\cdot\nabla_\eta\nabla_{\eta'}\cdot\mathbf{r}' \tag{0.11}
$$

The integrals in the Lagrangian (Eq. 0.5) which corresponds to integrating out rotational degrees of freedom are for the left-hand side

$$
\begin{aligned}
&\int_\Omega [\mathcal{O}(K)]^T[\mathcal{O}(K)] = \frac{1}{16\pi^2}\int_0^\pi\int_0^{2\pi}\int_0^\pi\int_0^{2\pi}[H^{\theta\phi\theta'\phi'}K]^T[H^{\theta\phi\theta'\phi'}K]\sin\theta\sin\theta' d\theta d\phi d\theta' d\phi' \\
=&\frac{1}{12\pi}\int_0^\pi\int_0^{2\pi}\int_0^\pi\int_0^{2\pi}\Big( \\
&[\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'K]^T[\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'K]\cos^2(\phi)\sin^2(\theta) \\
&+[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'K]^T[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'K]\sin^2(\phi)\sin^2(\theta) \\
&+[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'K]^T[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'K]\cos^2(\theta) \\
&+([\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'K]^T[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'K]+[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'K]^T[\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'K])\cos(\phi)\sin(\phi)\sin^2(\theta) \\
&+([\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'K]^T[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'K]+[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'K]^T[\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'K])\cos(\phi)\sin(\theta)\cos(\theta) \\
&+([\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'K]^T[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'K]+[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'K]^T[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'K])\sin(\phi)\sin(\theta)\cos(\theta)\Big) \\
&\sin\theta\sin\theta' d\theta d\phi d\theta' d\phi' \\
=&\int_0^\pi\int_0^{2\pi}\frac{1}{3}\Big([\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'K]^T[\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'K]+[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'K]^T[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'K] \\
&+[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'K]^T[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'K]\Big)\sin\theta' d\theta' d\phi' \\
=&\frac{1}{9}\sum_{\nu,\nu'\in x,y,z}\Big(\frac{\partial^2}{\partial\eta_\nu\partial\eta'_{\nu'}}K\Big)^T\Big(\frac{\partial^2}{\partial\eta_\nu\partial\eta'_{\nu'}}K\Big)
\end{aligned}
$$

and the right-hand side:

$$
\begin{aligned}
&\int_\Omega [\mathcal{O}(U)]^T[\mathcal{O}(K)] = \frac{1}{16\pi^2}\int_0^\pi\int_0^{2\pi}\int_0^\pi\int_0^{2\pi}[H^{\theta\phi\theta'\phi'}U]^T[H^{\theta\phi\theta'\phi'}K]\sin\theta\sin\theta' d\theta d\phi d\theta' d\phi' \\
=&\frac{1}{12\pi}\int_0^\pi\int_0^{2\pi}\int_0^\pi\int_0^{2\pi}\Big( \\
&[\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'U]^T[\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'K]\cos^2(\phi)\sin^2(\theta) \\
&+[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'U]^T[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'K]\sin^2(\phi)\sin^2(\theta) \\
&+[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'U]^T[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'K]\cos^2(\theta) \\
&+([\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'U]^T[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'K]+[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'U]^T[\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'K])\cos(\phi)\sin(\phi)\sin^2(\theta) \\
&+([\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'U]^T[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'K]+[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'U]^T[\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'K])\cos(\phi)\sin(\theta)\cos(\theta) \\
&+([\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'U]^T[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'K]+[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'U]^T[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'K])\sin(\phi)\sin(\theta)\cos(\theta)\Big) \\
&\sin\theta\sin\theta' d\theta d\phi d\theta' d\phi' \\
=&\int_0^\pi\int_0^{2\pi}\frac{1}{3}\Big([\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'U]^T[\frac{\partial}{\partial\eta_x}\nabla_{\eta'}\cdot\mathbf{r}'K]+[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'U]^T[\frac{\partial}{\partial\eta_y}\nabla_{\eta'}\cdot\mathbf{r}'K] \\
&+[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'U]^T[\frac{\partial}{\partial\eta_z}\nabla_{\eta'}\cdot\mathbf{r}'K]\Big)\sin\theta' d\theta' d\phi' \\
=&\frac{1}{9}\sum_{\nu,\nu'\in x,y,z}\Big(\frac{\partial^2}{\partial\eta_\nu\partial\eta'_{\nu'}}U\Big)^T\Big(\frac{\partial^2}{\partial\eta_\nu\partial\eta'_{\nu'}}K\Big)
\end{aligned}
$$

# References


[1] J. Serre, International Journal of Quantum Chemistry **26**, 593 (1984).

[2] M. Busch, M. D. Wodrich, and C. Corminboeuf, Chemical science **6**, 6754 (2015).

[3] M. Balmith, M. Faya, and M. E. Soliman, Chemical biology & drug design **89**, 297 (2017).

[4] N. Beerenwinkel, T. Sing, T. Lengauer, J. Rahnenführer, K. Roomp, I. Savenkov, R. Fischer, D. Hoffmann, J. Selbig, K. Korn, *et al.*, Bioinformatics **21**, 3943 (2005).

[5] K. El Hage, V. Pandyarajan, N. B. Phillips, B. J. Smith, J. G. Menting, J. Whittaker, M. C. Lawrence, M. Meuwly, and M. A. Weiss, Journal of Biological Chemistry **291**, 27023 (2016).

[6] G. Ceder, Y.-M. Chiang, D. Sadoway, M. Aydinol, Y.-I. Jang, and B. Huang, Nature **392**, 694 (1998).

[7] B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, and C. Corminboeuf, Chemical science **9**, 7069 (2018).

[8] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Phys. Rev. Lett. **117**, 135502 (2016).

[9] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).

[10] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[11] M. Head-Gordon, R. J. Rico, M. Oumi, and T. J. Lee, Chemical Physics Letters **219**, 21 (1994).

[12] G. D. Purvis III and R. J. Bartlett, The Journal of Chemical Physics **76**, 1910 (1982).

[13] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, Chemical Physics Letters **157**, 479 (1989).

[14] C. Møller and M. S. Plesset, Physical review **46**, 618 (1934).

[15] R. Krishnan and J. A. Pople, International Journal of Quantum Chemistry **14**, 91 (1978).

[16] D. Ceperley and B. Alder, Science **231**, 555 (1986).

[17] O. A. von Lilienfeld, Angewandte Chemie International Edition **57**, 4164 (2018).

[18] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, J. Chem. Theory Comput. **13**, 5255 (2017).

[19] F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld, J. Chem. Phys. **148**, 241717 (2018).

[20] A. S. Christensen, F. A. Faber, and O. A. von Lilienfeld, The Journal of Chemical Physics **150**, 064105 (2019).

[21] A. L. Samuel, IBM Journal of research and development **3**, 210 (1959).

[22] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, IEEE Trans. Neural Netw. **12**, 181 (2001).

[23] B. Schölkopf and A. J. Smola, "Learning with kernels: support vector machines, regularization, optimization, and beyond," (MIT press, 2002).

[24] V. Vovk, "Kernel ridge regression," in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, edited by B. Schölkopf, Z. Luo, and V. Vovk (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013) pp. 105–116.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: data mining, inference and prediction*, Springer series in statistics (Springer, New York, N.Y., 2001).

[26] S. De, A. P. Bartok, G. Csanyi, and M. Ceriotti, Phys. Chem. Chem. Phys. **18**, 13754 (2016).

[27] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, **3** (2017).

[28] A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, Phys. Rev. Lett. **120**, 036002 (2018).

[29] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).

[30] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, J. Chem. Theory Comput. **9**, 3404 (2013).

[31] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, J. Chem. Theory Comput. **11**, 2087 (2015).

[32] H. Huo and M. Rupp, arXiv preprint arXiv:1704.06439 (2017).

[33] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).

[34] A. N. Tikhonov and V. y. Arsenin, *Solutions of ill-posed problems* (Vh Winston, 1977).

[35] P. C. Hansen, *Discrete inverse problems: insight and algorithms*, Vol. 7 (Siam, 2010).

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, International Journal of Computer Vision (IJCV) **115**, 211 (2015).

[37] K. He, X. Zhang, S. Ren, and J. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 770–778.

[38] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, J. Chem. Phys. **148**, 241722 (2018).

[39] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017* (2017).

[40] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, in *Advances in Neural Information Processing Systems* (2015) pp. 2215–2223.

[41] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Nat. Commun. **8**, 13890 (2017).

[42] A. Kamath, R. A. Vargas-Hernández, R. V. Krems, T. Carrington Jr, and S. Manzhos, The Journal of chemical physics **148**, 241702 (2018).

[43] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, The Journal of chemical physics **148**, 241709 (2018).

[44] N. Lubbers, J. S. Smith, and K. Barros, The Journal of chemical physics **148**, 241715 (2018).

[45] T. S. Hy, S. Trivedi, H. Pan, B. M. Anderson, and R. Kondor, The Journal of chemical physics **148**, 241745 (2018).

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, in *Advances in neural information processing systems* (2012) pp. 1097–1105.

[47] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, in *Eleventh annual conference of the international speech communication association* (2010).

[48] D. E. Rumelhart, G. E. Hinton, R. J. Williams, *et al.*, Cognitive modeling **5**, 1 (1988).

[49] J. Kiefer, J. Wolfowitz, *et al.*, The Annals of Mathematical Statistics **23**, 462 (1952).

[50] D. C. Liu and J. Nocedal, Mathematical programming **45**, 503 (1989).

[51] D. Kingma and J. Ba, arXiv preprint arXiv:1412.6980 (2014).

[52] K. R. Müller, M. Finke, N. Murata, K. Schulten, and S. Amari, Neural Comp. **8**, 1085 (1996).

[53] C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik, and J. S. Denker, in *Advances in Neural Information Processing Systems* (1994) pp. 327–334.

[54] W. Dubitzky, M. Granzow, and D. P. Berrar, *Fundamentals of data mining in genomics and proteomics* (Springer Science & Business Media, 2007).

[55] T. Fushiki, Statistics and Computing **21**, 137 (2011).

[56] B. Huang and O. A. von Lilienfeld, J. Chem. Phys. **145**, 161102 (2016).

[57] D. Rogers and M. Hahn, J. Chem. Inf. Model. **50**, 742 (2010).

[58] T. D. Huan, A. Mannodi-Kanakkithodi, and R. Ramprasad, Phys. Rev. B **92**, 014106 (2015).

[59] A. Dalke, (2018).

[60] A. Glielmo, P. Sollich, and A. De Vita, Phys. Rev. B **95**, 214302 (2017).

[61] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Sci. Adv. **3**, e1603015 (2017).

[62] M. Gastegger, J. Behler, and P. Marquetand, Chem. Sci. **8**, 6924 (2017).

[63] R. Ramakrishnan, P. Dral, M. Rupp, and O. A. von Lilienfeld, Scientific Data **1**, 140022 (2014).

[64] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, J. Phys. Chem. Lett. **6**, 2326 (2015).

[65] C. R. Collins, G. J. Gordon, O. A. von Lilienfeld, and D. J. Yaron, The Journal of Chemical Physics **148**, 241718 (2018).

[66] B. Huang and O. A. von Lilienfeld, arXiv preprint arXiv:1707.04146 (2017), submitted to Nature.

[67] K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev, The Journal of Chemical Physics **148**, 241727 (2018).

[68] W. Pronobis, A. Tkatchenko, and K.-R. Müller, J. Chem. Theory Comput. (2018).

[69] O. T. Unke and M. Meuwly, The Journal of Chemical Physics **148**, 241708 (2018).

[70] B. Nebgen, N. Lubbers, J. S. Smith, A. E. Sifain, A. Lokhov, O. Isayev, A. E. Roitberg, K. Barros, and S. Tretiak, J. Chem. Theory Comput. (2018).

[71] M. Eickenberg, G. Exarchakis, M. Hirn, S. Mallat, and L. Thiry, The Journal of Chemical Physics **148**, 241732 (2018).

[72] J. Behler, J. Chem. Phys. **134**, 074106 (2011).

[73] B. G. Sumpter and D. W. Noid, Chem. Phys. Lett. **192**, 455 (1992).

[74] S. Lorenz, A. Gross, and M. Scheffler, Chem. Phys. Lett. **395**, 210 (2004).

[75] S. Manzhos and T. Carrington, Jr., J. Chem. Phys. **125**, 084109 (2006).

[76] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).

[77] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).

[78] S. Manzhos, R. Dawes, and T. Carrington, Int. J. Quantum Chem. **115**, 1012 (2015).

[79] V. Botu and R. Ramprasad, Int. J. Quantum Chem. **115**, 1074 (2015).

[80] M. Rupp, R. Ramakrishnan, and O. A. von Lilienfeld, J. Phys. Chem. Lett. **6**, 3309 (2015).

[81] J. S. Smith, O. Isayev, and A. E. Roitberg, Chem. Sci. **8**, 3192 (2017).

[82] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, New J. Phys. **15**, 095003 (2013).

[83] R. Ramakrishnan and O. A. von Lilienfeld, CHIMIA **69**, 182 (2015).

[84] A. E. Sifain, N. Lubbers, B. T. Nebgen, J. S. Smith, A. Y. Lokhov, O. Isayev, A. E. Roitberg, K. Barros, and S. Tretiak, ChemRxiv (2018), 10.26434/chemrxiv.6638981.v1.

[85] K. T. Schütt, M. Gastegger, A. Tkatchenko, and K.-R. Müller, arXiv preprint arXiv:1806.10349 (2018).

[86] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, **52**, 2864 (2012).

[87] F. Brockherde, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, Nature Communications **8**, 872 (2017).

[88] A. V. Sinitskiy and V. S. Pande, arXiv preprint arXiv:1809.02723 (2018).

[89] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, ACS Central Science **5**, 57 (2018).

[90] P. J. Stevens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, J. Phys. Chem. **98**, 11623 (1993).

[91] J. Sadowski and J. Gasteiger, Chemical Reviews **93**, 2567 (1993).

[92] R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, and E. L. Willighagen, Journal of chemical information and modeling **46**, 991 (2006).

[93] M. Ernzerhof and G. E. Scuseria, J. Chem. Phys. **110**, 5029 (1999).

[94] C. Adamo and V. Barone, J. Chem. Phys. **110**, 6158 (1999).

[95] L. A. Burns, J. C. Faver, Z. Zheng, M. S. Marshall, D. G. A. Smith, K. Vanommeslaeghe, A. D. MacKerellJr., K. M. MerzJr., and C. D. Sherrill, The Journal of Chemical Physics **147**, 161727 (2017).

[96] M. S. Marshall and C. D. Sherrill, Journal of Chemical Theory and Computation **7**, 3978 (2011).

[97] C. L. B. III and M. Karplus, The Journal of Chemical Physics **79**, 6312 (1983).

[98] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, Journal of Computational Chemistry **30**, 1545 (2009).

[99] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, The Journal of Chemical Physics **79**, 926 (1983).

[100] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, The Journal of Physical Chemistry B **102**, 3586 (1998).

[101] Stote, RH, States, DJ, and Karplus, M, J. Chim. Phys. **88**, 2419 (1991).

[102] P. J. Steinbach and B. R. Brooks, Journal of Computational Chemistry **15**, 667 (1994).

[103] S. Grimme, J. G. Brandenburg, C. Bannwarth, and A. Hansen, The Journal of Chemical Physics **143**, 054107 (2015).

[104] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, npj Computational Materials **1**, 15010 (2015).

[105] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).

[106] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, Acta Crystallographica Section B Structural Science **58**, 364 (2002).

[107] G. Bergerhoff, R. Hundt, R. Sievers, and I. D. Brown, Journal of Chemical Information and Computer Sciences **23**, 66 (1983).

[108] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, and C. Wolverton, Physical Review B **96**, 024104 (2017).

[109] R. Armiento *et al.*, *The High-Throughput Toolkit (httk)*, http://httk.openmaterialsdb.se/.

[110] P. E. Blöchl, Physical Review B **50**, 17953 (1994).

[111] G. Kresse and D. Joubert, Physical Review B **59**, 1758 (1999).

[112] G. Kresse and J. Furtmüller, *Vienna Ab Initio Simulation Package, Users Guide* (The University of Vienna, Vienna, 2007).

[113] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical Review Letters **77**, 3865 (1996).

[114] H. J. Monkhorst and J. D. Pack, Physical Review B **13**, 5188 (1976).

[115] K. Lejaeghere, V. Van Speybroeck, G. Van Oost, and S. Cottenier, Critical Reviews in Solid State and Materials Sciences **39**, 1 (2014).

[116] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[117] A. Tkatchenko and M. Scheffler, Phys. Rev. Lett. **102**, 073005 (2009).

[118] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, A. Tkatchenko, and K.-R. Müller, arXiv preprint arXiv:1706.08566 (2018).

[119] K. Burke, J. Chem. Phys. **136**, 150901 (2012).

[120] W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory* (Wiley-VCH, 2002).

[121] A. J. Cohen, P. Mori-Sánchez, and W. Yang, Chemical Reviews **112**, 289 (2012).

[122] R. E. Plata and D. A. Singleton, J. Am. Chem. Soc. **137**, 3811 (2015).

[123] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, and K. A. Lyssenko, Science **355**, 49 (2017).

[124] J. Barker, J. Bulin, J. Hamaekers, and S. Mathias, arXiv preprint arXiv:1611.05126 (2016).

[125] O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, and A. Knoll, Int. J. Quantum Chem. **115**, 1084 (2015).

[126] C. R. Collins, G. J. Gordon, O. A. von Lilienfeld, and D. J. Yaron, arXiv preprint arXiv:1701.06649 (2016).

[127] P. O. Dral, O. A. von Lilienfeld, and W. Thiel, Journal of Chemical Theory and Computation **11**, 2120 (2015).

[128] W. Weber and W. Thiel, Theor. Chem. Acc. **103**, 495 (2000).

[129] P. O. Dral, X. Wu, L. Spörkel, A. Koslowski, and W. Thiel, J. Chem. Theory **12**, 1097 (2016).

[130] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, J. Comput. Aided Mol. Des. **30**, 595 (2016).

[131] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, ICLR (2016).

[132] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, J. Chem. Phys. **106**, 1063 (1997).

[133] A. L. Hickey and C. N. Rowley, J. Phys. Chem. A **118**, 3678 (2014).

[134] Ralf Stowasser and Roald Hoffmann, J. Am. Chem. Soc. **121**, 3414 (1999).

[135] P. Sinha, S. E. Boesch, C. Gu, R. A. Wheeler, and A. K. Wilson, J. Phys. Chem. A **108**, 9213 (2004).

[136] R. C. Geary, Biometrika **27**, 310 (1935).

[137] D. F. DeTar, J. Phys. Chem. A **111**, 4464 (2007).

[138] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. G. III, and W. M. Skiff, J. Am. Chem. Soc. **114**, 10024 (1992).

[139] J. Besnard, G. F. Ruda, V. Setola, K. Abecassis, R. M. Rodriguiz, X.-P. Huang, S. Norval, M. F. Sassano, A. I. Shin, L. A. Webster, F. R. C. Simeons, L. Stojanovski, A. Prat, N. G. Seidah, D. B. Constam, G. R. Bickerton, K. D. Read, W. C. Wetsel, I. H. Gilbert, B. L. Roth, and A. L. Hopkins, Nature **492**, 215 (2012).

[140] E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Côté, B. K. Shoichet, and L. Urban, Nature **486**, 361 (2012).

[141] R. W. Huigens III, K. C. Morrison, R. W. Hicklin, T. A. Flood Jr, M. F. Richter, and P. J. Hergenrother, Nature chemistry **5**, 195 (2013).

[142] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors* (Wiley-VCH, Weinheim, 2009).

[143] J.-L. Faulon, D. P. Visco, Jr., and R. S. Pophale, J. Chem. Inf. Comp. Sci. **43**, 707 (2003).

[144] J. Visco, R. S. Pophale, M. D. Rintoul, and J. L. Faulon, J. Mol. Graph. Model. **20**, 429 (2002).

[145] R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. K. Wegner, and E. Willighagen, J. Chem. Inf. Model. **46**, 991 (2006).

[146] G. Landrum, http://www.rdkit.org **3**, 2012 (2014).

[147] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, arXiv preprint arXiv:1702.05532 (2017).

[148] Hoerl, E. Arthur, Kennard, and W. Robert, Technometrics , 80 (2000).

[149] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).

[150] R. M. Neal, *Bayesian Learning for Neural Networks* (Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996).

[151] H. Zou and T. Hastie, J. R. Stat. Soc. Series. B Stat. Methodol. **67**, 301 (2005).

[152] L. Breiman, Machine learning **45**, 5 (2001).

[153] T. Desautels, A. Krause, and J. W. Burdick, J. Mach. Learn. Res. **15**, 4053 (2014).

[154] "Google hypertune. https://cloud.google.com/ml/ (accessed 2016)," .

[155] D. P. Tew, W. Klopper, M. Heckert, and J. Gauss, J. Phys. Chem. A **111**, 11242 (2007).

[156] P. Kirkpatrick and C. Ellis, Nature **432**, 823 (2004).

[157] F. Jensen, *Introduction to Computational Chemistry* (John Wiley, West Sussex, England, 2007).

[158] K. Hansen, F. Biegler, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, J. Phys. Chem. Lett. **6**, 2326 (2015).

[159] O. A. von Lilienfeld, Int. J. Quantum Chem. **113**, 1676 (2013).

[160] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, Phys. Rev. B **89**, 205118 (2014).

[161] O. A. von Lilienfeld, *Many-Electron Approaches in Physics, Chemistry and Mathematics*, edited by V. Bach and L. D. Site, Mathematical Physics Studies, Vol. IX (Springer, 2014) p. 417.

[162] K. Y. S. Chang, S. Fias, R. Ramakrishnan, and O. A. von Lilienfeld, J. Chem. Phys. **144**, 174110 (2016).

[163] S. Mathias, Master thiesis (2015).

[164] A. P. Bartók and G. Csányi, International Journal of Quantum Chemistry **115**, 1051 (2015).

[165] B. M. Axilrod and E. Teller, J. Chem. Phys **11**, 299 (1943).

[166] Y. Muto, J. Phys. Math. Soc **17**, 629 (1943).

[167] A. S. Christensen, M. Elstner, and Q. Cui, The Journal of chemical physics **143**, 084123 (2015).

[168] G. Pilania, J. E. Gubernatis, and T. Lookman, Computational Materials Science **129**, 156 (2017).

[169] L. C. Blum and J.-L. Reymond, J. Am. Chem. Soc. **131**, 8732 (2009).

[170] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, The Journal of chemical physics **79**, 926 (1983).

[171] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. a. Swaminathan, and M. Karplus, Journal of computational chemistry **4**, 187 (1983).

[172] S. Grimme, J. G. Brandenburg, C. Bannwarth, and A. Hansen, The Journal of chemical physics **143**, 054107 (2015).

[173] A. P. Bartok, S. De, C. Poelking, N. Bernstein, J. Kermode, G. Csanyi, and M. Ceriotti, arXiv preprint arXiv:1706.00179 (2017).

[174] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Int. J. Quantum Chem. **115**, 1094 (2015).

[175] O. A. von Lilienfeld, J. Chem. Phys. **131**, 164102 (2009).

[176] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Scientific reports **3**, 2810 (2013).

[177] R. Ramakrishnan and O. A. von Lilienfeld, "Machine learning, quantum chemistry, and chemical space," in *Reviews in Computational Chemistry*, Vol. 30 (John Wiley & Sons, Inc., 2017) pp. 225–256.

[178] X. Gonze, Phys. Rev. A **52**, 1096 (1995).

[179] A. Putrino, D. Sebastiani, and M. Parrinello, J. Chem. Phys. **113**, 7102 (2000).

[180] R. G. Parr and W. Yang, *Density functional theory of atoms and molecules* (Oxford Science Publications, 1989).

[181] P. Geerlings, F. D. Proft, and W. Langenaeker, Chem. Rev. **103**, 1793 (2003).

[182] O. A. von Lilienfeld, R. Lins, and U. Rothlisberger, Phys. Rev. Lett. **95**, 153002 (2005).

[183] D. Sheppard, G. Henkelman, and O. A. von Lilienfeld, J. Chem. Phys. **133**, 084104 (2010).

[184] A. Solovyeva and O. A. von Lilienfeld, Phys. Chem. Chem. Phys. **18**, 31078 (2016).

[185] S. Fias, F. Heidar-Zadeh, P. Geerlings, and P. W. Ayers, Proc. Natl. Acad. Sci. USA **114**, 11633 (2017).

[186] K. S. Chang and O. A. von Lilienfeld, Physical Review Materials **2**, 073802 (2018).

[187] R. P. Feynman, Phys. Rev. **56**, 340 (1939).

[188] S. Manzhos, X. Wang, R. Dawes, and T. Carrington, J. Phys. Chem. A **110**, 5295 (2006).

[189] S. Manzhos and T. Carrington, J. Chem. Phys. **125**, 084109 (2006).

[190] S. Manzhos and T. Carrington, J. Chem. Phys. **129**, 224104 (2008).

[191] J. Cui and R. V. Krems, J. Phys. B **49**, 224001 (2016).

[192] G. Golub and C. Van Loan, *Matrix computations* (Johns Hopkins University Press, 1996).

[193] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning, www.GaussianProcess.org* (MIT Press, Cambridge, 2006) editor: T. Dietterich.

[194] B. M. Axilrod and E. Teller, J. Chem. Phys. **11**, 299 (1943).

[195] Y. Muto, J. Phys.-Math. Soc. Japan **17**, 629 (1943).

[196] J. Gasteiger and M. Marsili, Tetrahedron **36**, 3219 (1980).

[197] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, arXiv , arXiv:1802.09238 (2018).

[198] RDKit, online, "RDKit: Open-source cheminformatics," `http://www.rdkit.org`.

[199] J. S. Smith, O. Isayev, and A. E. Roitberg, Sci. Data **4**, 170193 (2017).

[200] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, "Gaussian09 Revision D.01," (2009).

[201] A. S. Christensen, F. A. Faber, B. Huang, L. A. Bratholm, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Qml: A python toolkit for quantum machine learning, https://github.com/qmlcode/qml," (2017).

[202] K. Madanakrishna, N. Edith, S. Vincent, van der Rest Guillaume, C. Duncan, and F. Gilles, Chem. Eur. J. **23**, 8414 (2017).

[203] N. Frank, Wiley. Interdiscip. Rev. Comput. Mol. Sci. **8**, e1327 (2017).

[204] P. Yang, F. P. Doty, M. A. Rodriguez, M. R. Sanchez, X. Zhou, and K. S. Shah, in *Symposium L – Nuclear Radiation Detection Materials – 2009*, MRS Online Proceedings Library, Vol. 1164 (2009).

[205] K. Biswas and M.-H. Du, Phys. Rev. B **86**, 014102 (2012).

[206] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, Phys. Rev. B **89**, 094104 (2014).

[207] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Physical review letters **114**, 105503 (2015).

[208] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, APL Materials **1**, 011002 (2013).

[209] S. Lany, Phys. Rev. B **78**, 245207 (2008).

[210] G. Hautier, S. Ong, A. Jain, C. Moore, and G. Ceder, Physical Review B - Condensed Matter and Materials Physics **85** (2012).

[211] A. E. Mattsson, R. Armiento, J. Paier, G. Kresse, J. M. Wills, and T. R. Mattsson, J. Chem. Phys. **128**, 084714 (2008).

[212] D. Pettifor, *Bonding and Structure of Molecules and Solids* (Oxford university press, 2002).

[213] A. R. Akbarzadeh, V. Ozoliņš, and C. Wolverton, Advanced Materials **19**, 3233 (2007).

[214] S. P. Ong, L. Wang, B. Kang, and G. Ceder, Chemistry of Materials **20**, 1798 (2008).

[215] O. Shyue Ping, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Computational Materials Science **68**, 314 (2013).

[216] "Pymatgen home page," (2014), http://pymatgen.org/.

[217] A. Togo and I. Tanaka, Scr. Mater. **108**, 1 (2015).

[218] W. Tang, E. Sanville, and G. Henkelman, Journal of Physics: Condensed Matter **21**, 084204 (2009).

[219] E. Sanville, S. D. Kenny, R. Smith, and G. Henkelman, Journal of Computational Chemistry **28**, 899 (2007).

[220] G. Henkelman, A. Arnaldsson, and H. Jònsson, Computational Materials Science **36**, 354 (2006).

[221] W. Marco and C. Röhr, Zeitschrift für Naturforschung B **62**, 1227–1234 (2014).

[222] C. Fonseca Guerra, J.-W. Handgraaf, E. J. Baerends, and F. M. Bickelhaupt, Journal of Computational Chemistry **25**, 189 (2004).

[223] R. Sarmiento-Pérez, T. F. Cerqueira, I. Valencia-Jaime, M. Amsler, S. Goedecker, S. Botti, M. A. Marques, and A. H. Romero, New Journal of Physics **15**, 115007 (2013).

[224] F. L. Hirshfeld, Theoretica chimica acta **44**, 129.

[225] B. F. Matthias, v. E. H. Nicolaas J. R., G. Célia Fonseca, and B. Evert Jan, Organometallics **15**, 2923 (1996).

[226] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, Journal of Physics: Condensed Matter **14**, 2745 (2002).